


3 1761 10374379 5



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743795>



Survey Methodology

A Journal of Statistics Canada

June 1992 Volume 18 Number 1





Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

June 1992 Volume 18 Number 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1992

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Chief, Author Services, Publications Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 1992

Price: Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	C. Patrick
	G.J.C. Hole	D. Roy
	F. Mayda (Production Manager)	M.P. Singh
	R. Platek (Past Chairman)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>U. of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	L.-P. Rivest, <i>Laval University</i>
J.-C. Deville, <i>INSEE</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
W.A. Fuller, <i>Iowa State University</i>	C.E. Särndal, <i>University of Montreal</i>
J.F. Gentleman, <i>Statistics Canada</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	C.M. Suchindran, <i>University of North Carolina</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat Inc.</i>
G. Kalton, <i>University of Michigan</i>	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

P. Lavallée, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 18, Number 1, June 1992

CONTENTS

In This Issue	1
Census Undercount Measurement Methods and Issues	
D.A. FREEDMAN and W.C. NAVIDI	
Should We Have Adjusted the U.S. Census of 1980?	3
Comment: S.E. FIENBERG	25
I.P. FELLEGI	29
N. CRESSIE	32
A.L. SCHIRM and S.H. PRESTON	35
J.A. HARTIGAN	44
T.P. SPEED	51
E.P. ERIKSEN and J.B. KADANE	52
Response from the Authors	59
N. CRESSIE	
REML Estimation in Empirical Bayes Smoothing of Census Undercount	75
G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI, L.K. SCHULTZ and J.H. TSAY	
Hierarchical and Empirical Bayes Methods for Adjustment of Census Undercount: The 1988 Missouri Dress Rehearsal Data	95
D. ROYCE	
A Comparison of Some Estimators of a Set of Population Totals	109
L. SWAIN, J.D. DREW, B. LAFRANCE and K. LANCE	
The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census	127
S.E. FIENBERG	
Bibliography on Capture-Recapture Modelling With Application to Census Undercount Adjustment	143
<hr/>	
L.K. ROE, J.F. CARLSON and D.A. SWANSON	
A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure	155
Z. XIA, P.S. LEVY, E.S.H. YU, Z. WANG and M. ZHANG	
Single Stage Cluster Sampling in Prevalence-Incidence Surveys: Some Issues Suggested by the Shanghai Survey of Alzheimer's Disease and Dementia	165

In This Issue

Despite the best efforts by statistical agencies in counting people in a census, a small undercount always remains. These undercounts are usually not uniformly distributed over various subgroups of the population and therefore they impact differently on various government programs that use census population figures. Consequently, methods of measuring undercounts, adjustment techniques, especially for local areas, and related issues have attracted a great deal of attention from policy makers and statisticians. The six articles included in the special section on **Census Undercount Measurement Methods and Issues** will be a valuable addition to the growing literature on this topic.

The first article in the section is a discussion paper by Freedman and Navidi. It reviews some of the statistical issues and arguments for and against adjusting the United States Census of 1980 as well as discusses statistical evidence presented in a trial against the Department of Commerce and the U.S. Bureau of the Census. The article is a continuation of the discussion between the authors and Ericksen, Kadane and Tukey who are proposing methodology for the adjustment. It also shows how some of the conflicting views were resolved by the trial court. The article is followed by very insightful and lively comments from several statisticians and a reply from the authors.

Cressie presents an empirical Bayes approach to prediction of undercount at subnational levels based on restricted maximum likelihood (REML). The claimed advantage of the REML estimators is that they do not tend to oversmooth the post-enumeration survey data as maximum likelihood estimation does. The REML estimators are compared with the maximum likelihood and method-of-moments estimators by simulation and example.

Prior to the 1990 U.S. Census, a dress rehearsal took place in the state of Missouri. Datta *et al.* use the data from this exercise to study procedures for modelling from census post-enumeration surveys. They consider both hierarchical and empirical Bayes approaches. The results indicate that both approaches lead to improvements on the dual system estimation approach. The authors conclude with an update in light of the adjustment of the actual 1990 U.S. Census.

Four estimators of the base population used as a benchmark in the Population Estimates Program of Statistics Canada are discussed by Royce. These are the unadjusted census counts, adjusted census counts, a preliminary test estimator and a composite estimator. The Weighted Mean Square Error is used as the basis for comparison of these estimators, not only for estimation of population totals but also for estimation of functions of population totals, such as population shares or growth rates *etc.*

Swain *et al.* give an overview of the Address Register that was created at Statistics Canada as a means of reducing undercoverage in the 1991 Census of Canada and represents a frame of the residential addresses for medium and large urban centres. Methodology, post-censal evaluation and future prospects are discussed.

The final article of the special section, by Fienberg, presents a selected annotated bibliography of the literature on capture-recapture estimation of population size. Capture-recapture estimation is the main method used to evaluate the completeness of the census counts and thus the article concentrates on literature related to the estimation of human populations.

Roe, Carlson and Swanson describe a variation of the Housing Unit Method to estimate the population of small rural areas. In this variation, local experts provide data about selected households. The estimates are compared to the census counts for three rural communities.

Xia *et al.* compare the statistical properties and costs of telescopic single stage cluster sampling with that of ordinary single stage cluster sampling. Telescopic single stage cluster sampling is an alternative when sub-sampling of clusters (*i.e.* two-stage cluster sampling) is not possible. The method has been used in the Shangai Survey of Alzheimer's Disease and Dementia, which serves as an illustration of how costs can be reduced without sacrificing precision.

The Editor

Should We Have Adjusted the U.S. Census of 1980?

D.A. FREEDMAN and W.C. NAVIDI¹

ABSTRACT

This paper reviews some of the arguments for and against adjusting the U.S. census of 1980, and the decision of the court.

KEY WORDS: Census; Adjustment; Post Enumeration Survey; Regression; Smoothing.

1. INTRODUCTION

Every ten years, the census gives a statistical portrait of the United States. Geographical detail makes these data unique. However, the counts have more than academic interest: they influence the distribution of power and money. The census is used to apportion Congress as well as local legislatures and to allocate tax money – \$40 billion per year in the late 1980s – to 39,000 state and local governments. For these purposes, the geographical distribution of the population matters, rather than counts for the nation as a whole. Indeed, the census is used as a basis for sharing out fixed resources: if one jurisdiction gets more, another must receive less. Adjusting the census is advisable only if the process brings us closer to a true picture of the distribution of the population.

A small undercount is thought to remain in the census, and this undercount is unlikely to be uniform. People who move at census time are hard to count; in rural areas, maps and address lists are incomplete. Central cities have heavy concentrations of poor and minority persons, who may be harder to enumerate. If the undercount can be estimated with good accuracy, especially at the local level, adjustments can – and should – be made to improve the census. Some statisticians argue that the undercount can be estimated well enough, others are skeptical: a bad adjustment may be worse than nothing.

Because of its resource implications, the undercount has attracted considerable attention in the media, the Congress, and the courts. After the 1980 census, New York City joined with other jurisdictions to sue the Department of Commerce, seeking to compel an adjustment based on demographic analysis and capture-recapture techniques. The Commerce Department resisted this pressure. The trial court framed the issue as follows:

“The plaintiffs contend that a statistical adjustment of the census will improve upon the accuracy of the census, thereby reducing the disproportionate undercount in the City and State [of New York]. The Census Bureau, however, contends that although the census counts are imperfect, a statistical adjustment of the census will inject even greater inaccuracies into the population count, and that therefore, a statistical adjustment of the census is not technically feasible or warranted at this time.” (674 F Supp 1091 = volume 674 of the Federal Supplement, page 1091).

¹ D.A. Freedman, Statistics Department, University of California, Berkeley, CA U.S.A. 94720; W.C. Navidi, Mathematics Department, University of Southern California, Los Angeles, CA U.S.A. 90089.

The 1980 case may seem dated, given that the census of 1990 has already been taken. However, among law suits that involve statistical principles, the 1980 census case was one of the most important and closely argued; there is still much to learn from it. This article will review some of the technical issues, and some of the findings of the court.

The balance of this section will sketch the background; for more details, see Cohen and Citro (1985) or Fay *et al.* (1988). There are two methods for evaluating the completeness of the counts in the U.S. Census: demographic analysis and capture-recapture. Demographic analysis uses administrative records (birth certificates, death certificates, immigration visas, *etc.*) to make independent estimates of population totals. The starting point is an accounting identity:

$$\text{Population} = \text{Births} - \text{Deaths} + \text{Immigration} - \text{Emigration}.$$

Demographic analysis provides estimates by age, sex and race but not ethnicity, because of gaps in the records. Data on immigration and emigration are incomplete; birth records are incomplete too, especially prior to 1935. Thus, the data going into the “identity” must be supplemented by a variety of imputations and adjustments. Furthermore, data on internal migration are lacking, so estimates are made primarily at the national level. This completes our sketch of demographic analysis.

Estimates of coverage for small areas (including states and cities) are based on capture-recapture techniques. Capture is in the census; recapture is in a sample survey conducted after the census. In 1980, there were two such surveys, or “*P*-samples:” the April and August CPS (Current Population Survey). Each record from the *P*-samples was matched against the census file to see if the corresponding person was “captured,” that is, counted in the census. Records that could not be matched indicated people who were missed by the census – or a failure in the matching process. These data were used to estimate the percentage of persons missed by the census, that is, the rate of omissions.

The census also had a small percentage of erroneous enumerations (for instance, people counted at two different addresses); the number was estimated by taking an “*E*-sample” of census records and trying to check them by field work. In effect, the net undercount was estimated by taking the difference between the omissions and erroneous enumerations. (For details, see Fay *et al.*, Chapter 5.) These undercount estimates were made as part of “PEP,” the Post Enumeration Program.

In 1980, there was a fair amount of missing data in the *P*- and *E*-samples: for instance, there was a 4% non-interview rate in the CPS; even after interview, a determination of match status could not be made for another 4% of the subjects. To see the effect of missing data, a variety of imputation schemes were considered, leading to 12 different series of PEP estimates for 66 subareas.

The 66 areas covered the whole U.S. They included cities like New York; states apart from these cities, like upstate New York; and whole states like Wyoming. A PEP “series” consists of 66 estimates, one for each study area; 9 of the 12 series were based on the April CPS, and 3 on the August CPS.

In the 1980 case, expert witnesses for plaintiffs included Gene Ericksen, Jay Kadane, and John Tukey. Their strategy for adjusting the census using PEP data was described in Ericksen and Kadane (1985). Freedman (among other statisticians and demographers) testified for the defendants, and Navidi was a consultant. A critique of the proposed adjustments was summarized in Freedman and Navidi (1986), to be referenced here as FN.

We now indicate some of the technical issues. According to experts from the Bureau of the Census:

- (a) There were substantial differences among the 12 PEP series, demonstrating that missing data were a serious problem.
- (b) The PEP estimates were subject to large biases, apart from the problems created by missing data.
- (c) Each PEP series was subject to unacceptably large sampling error.

Ericksen and Kadane responded that one of the PEP series ("PEP 2-9") was preferred, and that sampling error could be substantially reduced by regression modeling. They proposed a model with two equations. The first equation expresses the idea that y_i , the PEP estimate for study area i , is an unbiased estimate of the true undercount γ_i for that study area. Informally,

$$\text{PEP estimate for area } i = \text{True undercount in area } i + \text{Random error.}$$

Formally,

$$y_i = \gamma_i + \delta_i. \quad (1)$$

The second equation expresses a theory about the variation of the undercounts from area to area, in terms of a vector of explanatory variables X_i and a vector of hyper-parameters β . Informally,

$$\begin{array}{lcl} \text{True undercount} & \text{Linear combination of} & \\ \text{in area } i & \text{explanatory variables} & + \text{Random} \\ & \text{for area } i & \text{error.} \end{array}$$

Formally,

$$\gamma_i = X_i \cdot \beta + \epsilon_i. \quad (2)$$

The assumptions on the error terms can be stated as follows:

$$E(\delta_i) = E(\epsilon_i) = 0. \quad (3)$$

$$\text{var}\delta_i = K_i, \text{ var}\epsilon_i = \sigma^2. \quad (4)$$

$$\delta_1, \delta_2, \dots, \delta_{66}, \epsilon_1, \epsilon_2, \dots, \epsilon_{66} \text{ are independent.} \quad (5)$$

$$\delta_i \text{ and } \epsilon_i \text{ are normally distributed.} \quad (6)$$

In (4), K_i is the split-sample variance for y_i computed by the Bureau; randomness in K_i is ignored; σ^2 does not depend on i and is treated as constant even though it is estimated from the data. The role of assumptions, and departures from them, was examined in FN; also see the discussion papers and rejoinder, as well as sections 6-7 below.

The Ericksen-Kadane model was used in the 1980 case to smooth the PEP estimates, with the objective of reducing sampling error. The main focus of FN was a critique of that model. Ericksen, Kadane and Tukey (1989) – to be referenced here as EKT – replied to FN, and the present paper continues the exchange.

EKT cited a paper by Schirm and Preston (1987), which considers adjusting states and the District of Columbia by the "synthetic method." For instance, demographic analysis (with one set of assumptions on illegal immigration) estimated a national undercount rate of 5.9%

for blacks and 0.7% for whites in 1980. The synthetic method adjusts each state as follows: increase the number of blacks by 5.9% and the number of whites by 0.7%. In short, under-count rates are assumed to depend on race but not geographical area – or anything else.

This completes our summary of the technical background. For an update on the 1990 census, see Freedman (1991); some of the introductory material here was excerpted with minor changes from that paper. For other views, see Hogan and Wolter (1988), Schirm (1991), Wolter (1991), Wolter and Causey (1991), or Ericksen, Estrada, Tukey and Wolter (1991). The balance of the present paper responds to the salient points raised by EKT, and indicates how some of the the conflicting views were resolved by the trial court.

2. DO THE ADJUSTMENTS IMPROVE ON THE CENSUS?

The most important question is whether adjustments improve on the census counts. EKT “. . . are confident of improving upon the raw census count (p. 943)”’; indeed, there are “two simple [synthetic] adjustments that improve upon the census . . . the question of the Ericksen and Kadane model is not whether it proves that adjustment is feasible, but whether it improves upon the simpler methods (pp. 927-8) . . . Study of the method will not “prove” that an adjustment will improve the census. This has already been demonstrated by Schirm and Preston and the results of Tables 5 and 6 (p. 933).”

Thus, EKT’s Tables 5 and 6 are the main pieces of empirical evidence to show that adjustment will improve on the census. And Table 6 on erroneous enumerations is redundant, because the PEP estimates in Table 5 include the effect of erroneous enumerations. Table 5 is the critical one, and it is reproduced here for ease of reference. In our opinion, the table says very little about the possibility of improving on the census; to see why, some numerical detail is needed. (Schirm and Preston will be discussed in the next section.)

“Group 1” in the table consists of 16 central cities; “group 2” consists of other study areas that have relatively high minority populations; “group 3” consists of study areas with small minority populations. At best, the table shows that several methods for adjusting these groups are in general agreement. The table does not show that any of the methods improve on the accuracy of the census. It cannot, because there is no external standard against which to measure improvement.

Moreover, we believe the impression of agreement in the table to be largely illusory. There are dramatic differences among EKT’s preferred PEP series, or between these series and the synthetic adjustment of Schirm and Preston. Of course, drama depends on scale, and our next task is choosing units. Proponents of adjustment often use “loss functions” to make their argument; squared error is a common choice: see Ericksen, Estrada, Tukey and Wolter (1991, p. 20). EKT view Schirm and Preston as demonstrating census adjustment to be advantageous, so we compute the root mean square difference between the census and the “Synthetic B” line in Table 1, which is based on the Schirm and Preston adjustment. (The mean is weighted by population shares.)

$$\sqrt{.11 \times (.12)^2 + .44 \times (.06)^2 + .45 \times (.18)^2} \approx 0.13 \text{ of } 1\%.$$

In short,

$$\text{rms difference between census and synthetic } B = 0.13 \text{ of } 1\%. \quad (7)$$

Table 1

EKT's Table 5. Changes in National Population Shares Resulting When Counts are Adjusted by Sample Estimates Pooled Across Areas and Synthetic Estimates. [The entries for the three groups represent changes in shares, or differential undercounts; the entries in the last column represent total undercounts.]

PEP estimate	Group 1	Group 2	Group 3	Estimated national undercount rate
2-20	+ .52%	+ .09%	- .61%	+ 1.9%
3-20	+ .51%	+ .08%	- .59%	+ 1.7%
2-9	+ .50%	+ .06%	- .56%	+ 1.6%
3-9	+ .49%	+ .04%	- .53%	+ 1.4%
2-8	+ .41%	+ .04%	- .45%	+ 1.1%
3-8	+ .39%	+ .03%	- .42%	+ 1.0%
5-9	+ .31%	+ .25%	- .56%	+ 2.1%
5-8	+ .22%	+ .23%	- .45%	+ 1.7%
14-20	+ .21%	+ .02%	- .23%	- .2%
10-8	+ .19%	+ .07%	- .26%	+ .3%
14-9	+ .19%	- .01%	- .18%	- .5%
14-8	+ .10%	- .03%	- .07%	- 1.0%
Synthetic A	+ .17%	+ .14%	- .31%	+ 1.4%
Synthetic B	+ .12%	+ .06%	- .18%	+ 1.4%
Shares of Census Count	10.76%	44.24%	45.00%	

Notes: (i) Group 1 includes 16 central cities. Group 2 includes three state remainders (California, Maryland, and Texas, excluding Group 1 cities) and 17 whole states. All areas are at least 10% Black or Hispanic. Group 3 includes nine state remainders and 21 whole states. All Group 3 areas are less than 10% Black or Hispanic. (ii) The Synthetic A estimates assume that (a) Blacks have the same undercount rates as Hispanics, 5.9%; (b) the undercount rate of persons neither Black nor Hispanic is 0.3%; (c) the undercount rates for Blacks, Hispanics, and all others are invariant across geographic areas; and (d) there are 3 million undocumented aliens, 9.6% of whom are Black. (iii) Following Schirm and Preston (1987), the Synthetic B estimates assume that (a) the Black undercount rate is 5.9%; (b) Hispanics and other non-Blacks have an undercount rate of .7%; (c) the undercount rates for Blacks, Hispanics, and all others are invariant across geographic areas; and (d) there are 3 million undocumented aliens, 9.6% of whom are Black.

EKT prefer the first 8 of the PEP series (pp. 933 and 938). We next compute the rms difference between PEP 2-20 and 3-8, which are among EKT's preferred series. (PEP 2-20 and 3-8 were both based on the April CPS; differences between them are due only to procedures for handling missing data.)

rms difference between PEP 2-20 and 3-8 = 0.14 of 1%. (8)

EKT also recommend averaging as a way of eliminating indeterminacies (pp. 931 and 937). Table 2 compares population shares from the census, the synthetic B estimates, and the average preferred PEP estimates. We take the rms difference between the average preferred PEP and synthetic B:

rms difference between average preferred PEP and synthetic B = 0.25 of 1%. (9)

Table 2
Population Shares from the Census, the Synthetic B Estimates,
and the Average of EKT's Eight Preferred PEP Series
(2-20, 3-20, 2-9, 3-9, 2-8, 3-8, 5-9, 5-8).

	Group 1	Group 2	Group 3	Total
Average Preferred PEP – Synthetic B	.30%	.40%	– .34%	.00%
Census – Synthetic B	– .12%	– .06%	+ .18%	.00%
Average Preferred PEP	11.18%	44.34%	44.48%	100.00%
Synthetic B	10.88%	44.30%	44.82%	100.00%
Census	10.76%	44.24%	45.00%	100.00%

- A comparison of (7), (8) and (9) reveals three salient points:
- (a) the difference between the census and synthetic B is rather small;
 - (b) the range in the preferred PEP series is larger than the difference between the census and synthetic B;
 - (c) the difference between the average preferred PEP and synthetic B is twice the difference between the census and synthetic B.

EKT must view a difference of 0.13% as serious: see (7). On this scale, the PEP series do not agree among themselves. Furthermore, the PEP series are very different from the synthetic adjustment. Of course, the reason may be that Schirm and Preston did not go far enough. However, a National Academy of Sciences review panel – with Jay Kadane as a prominent member – reached the tentative conclusion that Schirm and Preston already over-adjusted the census: see Cohen and Citro (1985, p. 287).

The PEP estimates are in better agreement with the “synthetic A” adjustment in Table 1. But this is circular: the undercount rate for hispanics in synthetic A was estimated from PEP, while synthetic B was based on demographic analysis. Differences among the PEP estimates are an awkward reality; and so are differences between the PEP estimates and synthetic adjustments.

We now quote the principal claim made by EKT (p. 927):

“Our conclusion is that regardless of whether we use one of the simple methods or the composite method and regardless of how we vary the assumptions of the composite method, an adjustment reliably reduces population shares in states with few minorities and increases the shares of large cities.”

Giving more money to cities by changing the census counts is a good idea only if the adjustment reliably improves the accuracy of the census. Accuracy is the crucial issue, and we wish EKT would address it more directly. Their Table 5 is almost irrelevant.

3. SCHIRM AND PRESTON

Can synthetic adjustment reliably improve on the accuracy of the census? EKT think so, citing Schirm and Preston (1987) for the evidence. Schirm and Preston present two major arguments, one analytic and one based on simulation. However, both have serious flaws.

Table 3
A Counter-example to the Analytical Argument.
There are Two States and Two Races.

	White		Black		Total	
	Census count	True count	Census count	True count	Census count	True count
State A	90	89	1	2	91	91
State B	910	890	99	119	1,009	1,009
Total	1,000	979	100	121	1,100	1,100

The analytic argument (p. 966):

“Our finding is that synthetic adjustment will always move the estimated ratio of a state’s population to the national population closer to the true ratio if:

- (a) the state’s black undercount is closer to the national black undercount than it is to the national undercount for both races combined and
- (b) the state’s white undercount is closer to the national white undercount than it is to the national undercount for both races combined.”

As a matter of mathematics, this proposition is wrong. A counter-example is given in Table 3: state A, for instance, has by construction 89 whites and a census count of 90.

The counter-example has been set up to make the arithmetic easy; more complicated and realistic examples could undoubtedly be provided. In Table 3, the overall error in the census (white plus black) is 0, for each state and for the nation. Thus, the census gets the state shares right, and any adjustment will make matters worse. Error rates (with the true population as base) are shown in Table 4: Schirm and Preston’s conditions are satisfied. Synthetic adjustment moves both states farther from truth, as shown in Table 5; state B is helped, state A is hurt. To compute Table 5 from Table 3, the number of whites in state A is multiplied by:

$$\text{true national total for whites/national census total} = 979/1,000.$$

(10)

The arithmetic for the other cells is similar.

The counter-example may be informative, as a parable: state A is sparsely populated, with a small minority population; state B is heavily populated, and has a large, hard-to-count minority population. Synthetic adjustment may favor states of type B at the expense of type A. The mathematical error in Schirm and Preston’s appendix appears to be in their reasoning from display A.2. Professor Preston informs us (personal communication) that the theorem holds, with a more complicated set of conditions involving weighted averages.

Table 4
Undercounts from Table 3, in Percent.
(Negative undercounts correspond to overcounts.)

	White	Black	Total
State A	− 1.1%	50%	0%
State B	− 2.2%	17%	0%
Total	− 2.1%	17%	0%

Table 5
The Synthetic Adjustment, “Syn”.

	White		Black		Total	
	Syn	True count	Syn	True count	Syn	True count
State A	88	89	1	2	89	91
State B	891	890	120	119	1,011	1,009
Total	979	979	121	121	1,100	1,100

This completes our discussion of the analytic reasoning in Schirm and Preston. What about the simulation results? Basically, Schirm and Preston consider 51 areas (the states and D.C.) and two races (black and white). They set up a joint distribution for an assumed “true” population and the census counts; both are taken as stochastic. The census counts can be adjusted by the synthetic method, and the question is whether the raw counts or the adjusted counts are closer to the assumed true counts. Schirm and Preston actually consider several joint distributions, defined by different “scenarios,” that is, choices of parameters; the results are quite similar across scenarios. They also consider several loss functions, or measures of closeness.

- We focus on Scenario I, and make two brief comments.
- (a) The claimed improvement is rather modest. For example, on average, just over half the population lives in states whose shares are made more accurate by adjustment – no matter how small the improvement.
 - (b) The “true” population was constructed on the basis of the synthetic assumption – no systematic variation in undercount rates within race across geography; random variation was allowed. See equation (2) in Schirm and Preston. Thus, the definition of “truth” favors synthetic adjustment.

On the whole, however, Schirm and Preston have a reasonable argument. If the assumptions of the synthetic method more or less hold, its estimates will be good. There remains the crucial question: do those assumptions hold? what kind of geographical variation is there in undercount rates? On this score, Schirm and Preston offer no evidence. In the 1980 case, the trial court found that “the synthetic method simply ignores geographical variations and assumes that a person is as likely to be missed in the census whether he lives in Alabama or in Alaska. However, as defendants’ experts persuasively explained, this assumption that the undercount rates for the various age, race, and sex groups are constant from one subnational area to another has no basis in fact whatsoever . . . the synthetic method is simply inadequate as a means of adjusting the census.” (674 F Supp 1098, footnotes and citations omitted).

4. ADJUSTING SMALL AREAS

Statistical adjustment of census counts is more likely to be beneficial at fairly high levels of geographical aggregation (for instance, census regions or divisions). However, there are 39,000 state and local governments in the U.S., all claimants for tax money. Many of these jurisdictions are further subdivided, into city council seats, *etc*. If census counts are to be adjusted, they must for legal and policy reasons be adjusted at quite fine levels of geographical detail. Indeed, the proposal for 1990 is to adjust down to the block level. (A “block” is the smallest unit of census geography; there are 6.5 million blocks in the U.S.).

EKT discuss two synthetic methods for adjusting subareas of the 66 study areas, as well as a regression method (p. 941). In the end, however, there is no evidence that adjustment of small areas will improve on the raw census counts. With respect to 1980, EKT say (p. 943):

“For the 66 areas included in our study, we are confident of improving upon the raw census count, especially in those areas with large undercounts or overcounts where an adjustment is most needed. Our findings do not permit definitive conclusions for suburban areas, for central cities other than the 16 included in our data set, or for other rural or urban parts of individual states. To compute estimates for such areas, we would prefer not to extrapolate from the regression equations presented in this article.”

EKT go on to describe alternative designs for capture-recapture sampling, leaving open the question of small-area adjustment for 1990. Much of the dispute in 1980 centered on the feasibility of adjusting small sub-areas of the 66 study areas. To win its case, New York had to show such adjustments would improve on the census. EKT now seem to concede there was little evidence on this score.

5. AVERAGING AND SENSITIVITY ANALYSIS

The 12 PEP series were the results of a sensitivity analysis on missing data. Since the amount of missing data was large relative to the undercount, methods for handling missing data have impact. In response, EKT offer quite a variety of procedures for adjusting the census on the basis of the various PEP series, including: (a) eliminating discrepant series (pp. 937-9); (b) eliminating systematic differences between the series (pp. 937-8); (c) regression on other variables (the “composite” estimator, pp. 933ff); (d) averaging (pp. 931 and 937).

This list makes clear the essential indeterminacy of census adjustment schemes. And in this context, the use of averages to reduce indeterminacy needs discussion. Arbitrary modeling decisions may be defensible if they do not matter – the usual robustness argument. Sensitivity analysis (changing the assumptions to see if the results change) may refute the robustness argument. However, averaging the results from a sensitivity analysis is self-defeating. The different PEP series are not repeated measurements of the undercount. It is the spread in the PEP series that is interesting, not the average—because it is the spread (among, say, the April series) that demonstrates the impact of different modeling assumptions on the same data.

6. ASSUMPTIONS

EKT (p. 937) say the model improves on the PEP estimates and the synthetic method. The model does improve on the PEP estimates, if you grant its assumptions—equations (1) through (6) above. So far, however, these equations still seem quite implausible. Likewise, the model improves on the synthetic estimates only if it uses the additional variables in a sensible way, bringing us right back to assumptions.

At times, EKT seem to argue that the model can be inferred from the data (pp. 933ff). Of course, there is more to a regression model than choice of variables on the left hand side and the right hand side, although that is difficult enough, as will be seen below. There are many questions to answer: Why are effects linear and additive – equations (1) and (2) above? What about the assumptions on the errors – equations (3) through (6)? And so forth. EKT put forward no evidence to justify their assumptions, except by attempting to rebut our rebuttal (p. 931). Do they think a model is right unless it can be proved wrong?

In any case, we stand by our critique. For some data on correlation bias, see Fay *et al.* (1988, esp. sec. 6F); for a critique of Ericksen and Kadane's estimates, see Fellegi (1985, p.118). Other sources of bias in the PEP series include matching errors and errors in census-day address reports.

EKT argue that PEP is "conservative" (p. 931). This seems to be both wrong and irrelevant; wrong because the biases generally increase the apparent undercount: and irrelevant because geographical variation in the biases matters a great deal. Assumption (3) is rather unlikely: the errors probably do not have mean 0. The undercounts estimated by PEP are likely to be biased upward, the size of the bias depending on the area. For a review of the evidence, see Fay *et al.*, chap. 6; also see FN. The trial court in the 1980 case concluded:

"The evidence at trial established that the PEP was plagued by various errors caused by inadequacies in the PEP methodology. This type of error is referred to as 'bias.' A significant source of bias in the PEP arises because the process of matching people from the CPS to the census . . . is an extraordinarily difficult and inexact task. Because of inaccurate, irregular, and incomplete information in both the CPS and the census, the Bureau undoubtedly and inevitably made many errors in determining the match status of individuals enumerated in the CPS, thereby distorting the *P*-sample's undercount estimate. Moreover, the evidence at trial established that most of this matching error occurred because the Bureau erroneously determined many cases to be misses when they were in fact matches. This error, therefore, resulted in the PEP overstating the undercount. The extent of this error and the degree to which it varies from one geographic area to another is unknown." (674 F Supp 1100, footnotes and citations omitted).

We turn now to equations (4) and (5). Take the independence assumption. In 1980, there were 3 processing offices and 12 regional offices. EKT's counter: there were 400 district offices. Granted. There were also several dozen area managers, several hundred thousand census staff and about 1,500 CPS interviewers. The sources of error are numerous, and dependence seems likely. Processing offices, regional offices, managers, census interviewers, and CPS interviewers all must contribute components of error, to say nothing of respondents. Likewise, the constancy of σ^2 in (4) seems unlikely: different parts of the country are undercounted for different reasons, not readily captured in a linear regression equation.

We pointed out that random events like snowstorms might cause correlated errors in several areas; EKT respond that there were no snowstorms. This issue goes to the foundations of statistics: if the weather is good, the errors are independent; but in foul weather, all bets are off. The distributions in the model, and the statistical inferences, are therefore conditional on certain events. Which ones, and why?

Fortunately, we do not need to resolve the problem of conditional vs. unconditional inference. There was a major event that disrupted census operations over several states in the Pacific Northwest. Mt. St. Helens erupted in May 1980, while follow-up interviewing was in full swing.

7. OTHER ISSUES

7.1 Does it Matter which Series is Used?

At the level of precision EKT demand of the census, the different PEP series – even among their preferred ones – really do lead to quite different adjustments, as shown by equations (7) through (9). EKT, however, claim that the preferred PEP series all lead to similar adjustments. And to support their position they offer Table 11, which suggests for example that New York City has a differential undercount of 3.27% with an uncertainty of 0.62%.

For many purposes, a uniform undercount would not be material; it is differential undercounts that create inequities. The “area effects” seem to be measures of differential undercount – the policy variable of main interest.

The “area effects” in the table were computed by EKT as follows:

- (i) Restrict attention to 8 of the 12 PEP series.
- (ii) Smooth each of these using the regression model.
- (iii) For each area, take the average of the 8 estimates.
- (iv) Subtract the corresponding national estimate of undercount.

Table 6 below compares “area effects” with differences in the PEP estimates, attention being restricted to the preferred series based on the April CPS. Differences among these PEP estimates are due only to differences in the handling of missing data. Taking the range seems fair: reasons for data to be missing can differ from area to area, and so will the appropriate imputation procedure. Adding in the August series would increase the range, but some of the difference would be due to sampling error.

The table shows that for some areas, the effects are large relative to differences between PEP series, suggesting that missing data have little impact on the results. Upstate New York is an example. But for other areas, like Chicago, the reverse holds and imputation procedures matter.

All 66 areas are plotted in Figure 1. The *x*-axis shows the area effect; the *y*-axis shows the range in the preferred April PEP series. In root mean square (across the 66 areas), the spread among EKT’s preferred PEP series – based on the April CPS – is about 75% of the area effect. In other words, the impact of missing data (never mind other biases in PEP) is similar in magnitude to the effect EKT are trying to measure. Bringing in alternative imputation models would make matters even worse. Nor is averaging the results a good fix, for reasons given earlier.

Table 6
Comparing Area Effects with Differences in the PEP Estimates,
Restricted to Preferred Series Based on the April CPS.
Subareas Match those used in FN.

	Preferred April PEP series			Area effect
	Min.	Max.	Range	
Alabama	– .37	.60	.97	– 1.07
Alaska	2.79	3.53	.74	1.63
Los Angeles	4.56	7.72	3.16	3.16
San Diego	– .98	1.45	2.43	.65
San Francisco	4.31	6.25	1.94	2.31
Rest of California	2.84	3.92	1.08	1.03
Chicago	3.57	6.56	2.99	1.77
Rest of Illinois	1.21	1.75	.54	– 1.04
New York City	6.04	7.90	1.86	3.27
Rest of New York	– 1.61	– 1.44	.17	– 2.55
Wyoming	3.91	4.04	.13	1.16

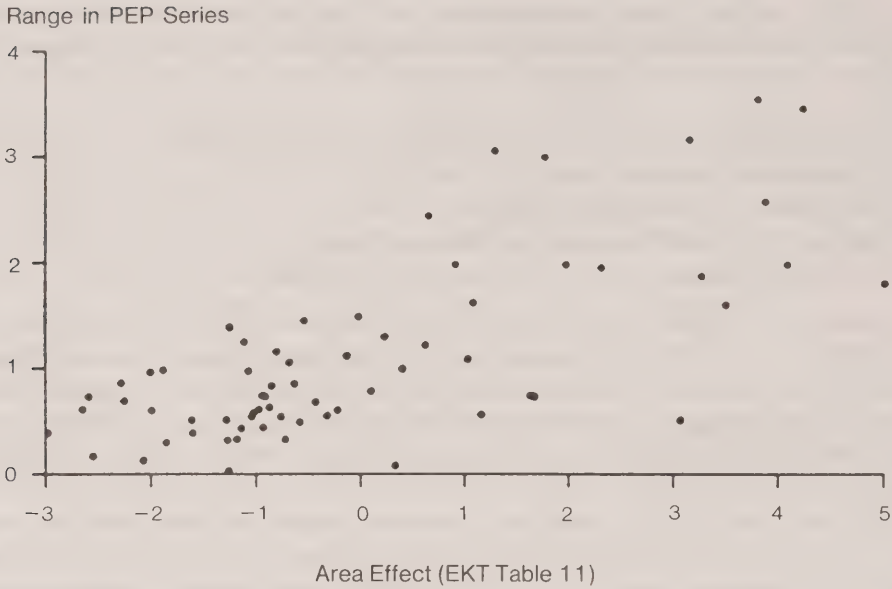


Figure 1. PEP and data quality. For each of the 66 study areas, the horizontal axis shows the EKT “area effect.” The vertical axis shows the range in the preferred April PEP series.

The positive association in Figure 1 is quite striking, and so is the change in the joint distribution when the area effect changes from negative to positive. Our explanation: PEP estimates of undercount are indicators of poor data quality – in PEP as well as the census. Large apparent undercounts indicate areas with poor data. In such areas, there is a lot of missing data, so the effect of changing the imputation rules will be large too. Areas that are hard to count are also hard to adjust. See FN p. 9 or Wolter (1986, p. 26, points 8 and 9).

There may be some reasonable way of choosing a compromise version among the PEP series. But why are any of the PEP series, or their averages, an improvement over the census? That is the crucial question, and EKT do not answer it. In our view, adjustment – whether by a synthetic method, or a PEP series, or a regression model, or any convex combination – will in the end be driven mainly by assumptions.

7.2 Which PEP Series is Best, and which Explanatory Variables should be Used?

At trial, and in their discussion of FN, Ericksen and Kadane recommended an adjustment based on PEP 2-9, apparently the most preferred of all 12 series. We chose PEP 10-8 as an alternative for study. EKT defend 2-9, and try to exclude 4 of the series – especially our foil 10-8. The arguments were reviewed in court and in FN (p. 8, the discussion, and the rejoinder p. 36). Our opinion remains the same: there is no rational basis for choosing 2-9 over 10-8.

EKT impute to us the position that “proportion urban” should have been considered as an independent variable (p. 934). This is not quite right. We felt that EKT’s choice of independent variables was somewhat arbitrary, and wanted to show that changing variables made a real

difference to the results – another sensitivity analysis. The difference was observed mainly for small areas (FN, p. 9). Since EKT no longer advocate adjustment of small areas in 1980, this argument may be moot.

There is one new twist to the reasoning: EKT argue for choosing models by “reliance on statistical criteria (p. 941).” In essence, they recommend choosing variables so as to minimize the rms residual in an OLS fit. However, the rms residual measures association in the data not correctness of underlying theory.

For reasons that remain unclear, EKT restrict attention to models with 2, 3, or 4 variables; and they require coefficients to have *t*-statistics of 2 or more. Their preferred equation seems to be:

$$\begin{aligned} \text{PEP 2-9} = & -2.23 + .079 \text{ min} + .036 \text{ crime} + .028 \text{ conv} + \text{residual} & (11) \\ & (-4.0) \quad (5.4) \quad (3.6) \quad (3.5) \\ & \text{rms residual} = 1.53. \end{aligned}$$

The right hand side variables are the percent minority in the study area, the crime rate, and the percent conventionally enumerated; *t*-statistics are shown in parentheses; the rms residual is computed using the unbiased divisor $n - p$. This equation is used only for variable selection; after the variables are chosen, the model is refitted by GLS: see (1-6) above, and FN for discussion.

The statistical logic is not apparent, and EKT's criteria have to be read quite literally. For example, here is another candidate equation:

$$\begin{aligned} \text{PEP 2-9} = & .120 \text{ min} + .026 \text{ crime} + .029 \text{ conv} - .176 \text{ pov} + \text{residual} & (12) \\ & (7.6) \quad (3.4) \quad (3.8) \quad (-4.4) \\ & \text{rms residual} = 1.49. \end{aligned}$$

The additional variable is the percentage of persons in the study area with incomes below the poverty level; the intercept was suppressed because the *t*-statistic was small. Equation (12) fits a little better than (11) in terms of rms residual, and “shows” that the undercount goes down as the percentage of poor people goes up – other things being equal. EKT reject this equation because the coefficient of “pov” is significantly negative rather than significantly positive.

Preconceptions about the undercount may be incompatible with the data, and best-subsets OLS may not be a suitable analytic technique. We reject neither interpretation, but our main conclusion is this. In the present context there are no objective, statistically defensible criteria for model selection. Much rides on the subjective judgment of the modeler.

With this in mind, we return to the points at issue – choosing a PEP series, and deciding between the crime rate or the percent urban as explanatory variables. As far as we can see, on the criteria chosen by EKT, the difference between crime rate and percent urban is trivial. And PEP 10-8 is clearly better than 2-9. See Table 7.

On pages 935 and 940 of EKT, σ denotes the rms residual. There is some conflict in notation, because we wrote σ^2 for $\text{Var}(\epsilon)$ in equations (2) and (4), following Ericksen and Kadane (1985, p. 105) or FN (p. 5). To avoid conflict, let $\text{SE}(\epsilon)$ be the estimated value for our σ ; this is what controls the standard errors of the 66 area undercounts computed by the Ericksen-Kadane model, as shown by equations (8) and (10) in FN. For PEP 10-8, the estimated $\text{SE}(\epsilon)$ is virtually 0, so a model based on 10-8 fits extremely well and the 66 area undercounts are very precisely estimated (Table 8).

Table 7
RMS Residuals from Regression Equations for PEP 2-9 and PEP 10-8.
Explanatory Variables Include Percent Minority,
Percent Conventionally Enumerated,
and Either the Crime Rate or the Percent Urban.

	Crime rate	Percent urban
PEP 2-9	1.53	1.54
PEP 10-8	1.35	1.33

Table 8
SE (ϵ) and the RMS for the 66 Study Areas; PEP 2-9 and PEP 10-8.
The Models Include Percent Minority, Percent Conventionally Enumerated,
and Either the Crime Rate or the Percent Urban.

	Crime rate		Percent urban	
	SE (ϵ)	rms area SE	SE (ϵ)	rms area SE
PEP 2-9	.75	.65	.76	.65
PEP 10-8	.00	.28	.00	.25

Notes: Let K be a 66×66 diagonal matrix, whose (i,i) element is K_i . Let X be the 66×4 matrix of explanatory variables. Let $H = X(X^T X)^{-1} X^T$ and $\Gamma^{-1} = K^{-1} + \text{SE}(\epsilon)^{-2} (I - H)$. The 66 area undercounts are estimated by the Ericksen-Kadane model as $\Gamma K^{-1} y$, where y is the 66×1 vector of PEP estimates. The rms SE for the 66 study areas is $\sqrt{\text{trace } \Gamma / 66}$. For details, see FN. At trial, Ericksen and Kadane estimated SE (ϵ) from 51 study areas (whole states and DC); we followed suit in FN. Here, we use the 66 study areas, since that seems to be EKT's current recommendation. The difference is noticeable.

On “statistical criteria,” contrary to the claims made by EKT, 10-8 is preferred to 2-9 and percent urban is just as good an explanatory variable as the crime rate. Their qualitative critique seems off the mark too. Of course, different urban areas are different, just as EKT say. So are different central cities. Similarly, minority persons living in central cities are likely to be different from those in suburbs. And so forth. All of EKT’s variables are “blurred predictors” of undercount, and some are blurrier than the percent urban (p. 934).

With respect to this set of issues, the judge in the 1980 case was harder on Ericksen and Kadane than we are:

“Moreover, as defendants’ experts persuasively explained, no one series of PEP estimates can be reliably shown to be superior to the others, or indeed, to the census itself, because there is insufficient knowledge with respect to which PEP procedures are better suited for measuring census undercount. While two of plaintiffs’ experts expressed a preference for the ‘series 2-9’ PEP estimates based upon the hypothesis that the PEP procedures employed in arriving at those estimates were superior to the procedures used for the other PEP estimates, the plaintiffs’ experts offered nothing more than unsupported assumptions in support of that position. On the other hand, the defendants’ experts offered equally plausible assumptions which favored different PEP procedures, producing dramatically different PEP estimates.” (674 F Supp 1102, footnotes and citations omitted.)

7.3 Simulation Studies

We had a simulation study making three points: (a) you could not infer from the data which variables go into the model, (b) standard errors depend on assumptions about disturbance terms, and (c) the standard errors computed by Ericksen and Kadane were quite optimistic. We had two additional points on this topic: (d) standard errors do not measure the impact of bias; (e) the Ericksen-Kadane smoothing simply passes through any bias in PEP that is well related to the explanatory variables.

Points (a) through (e) are real obstacles to showing that the model improves on the PEP estimates. EKT do not comment on points (b), (d) and (e). They deny (a), but more or less concede point (c). For our part, we concede that in our simulation – which grants half the model – regression does reduce sampling error. We still think (a) is right, as will be argued below. And in other contexts, smoothing may actually increase sampling error (Ylvisaker 1991 p. 7).

EKT (p. 943) criticize our study, because it covered only models with three variables in the equation and did not restrict the *t*-statistics. So we repeat the simulation here. In essence, we take PEP 10-8 as “truth,” and add for each of the 66 study areas *i* a random error with variance *K_i*, as in (4). This grants equation (1) and the assumptions on δ_i . We choose variables according to the procedure outlined by EKT (p. 935), and fit the regression model, repeating the whole process 100 times.

Table 9 shows the variables selected in the first 10 runs. As will be seen, there is no consistency – except that the percentage “conventionally enumerated” always comes in. Over the 100 runs – excluding the ones that produced no acceptable model – the nominal rms error was about 30% too small, and improvement of the composite estimator over PEP was exaggerated by a factor of 1.75. Assumptions matter.

Table 9
A Simulation Experiment on Variable Selection; PEP 10-8 is Taken as “Truth.”

Run	CC	Min	Crime	Conv	Ed	Pov	Lang	MU
1			x	x	x			
2		x		x				
3		x		x				
4	x		x	x				
5	x			x				
6		x		x				
7			x	x	x			
8		x		x				
9	x			x				
10	There was no model satisfying EKT’s criteria							

Notes: CC is an indicator for central cities; Min, the percentage of minorities; Crime, the crime rate; Conv, the percentage who were conventionally enumerated; Ed, the percentage with no high school degree; Pov, the percentage below the poverty line; Lang, the percentage who have difficulty with English; MU, the percentage living in multiple-unit housing.

Table 10
A Simulation Experiment on Variable Selection.
PEP 2-9 is Taken as “Truth”; Percent Urban (Urb) is Permitted as an Explanatory Variable.
The Table Shows The Number of time Each Variable is Entered, and The Average
of its Coefficient (Over The Time it Enters); 100 Data Sets were Generated.

Variable	No. of times entered	Average coefficient
CC	17	2.954
Min	82	0.071
Crime	53	0.053
Conv	93	0.028
Ed	5	0.085
Pov	1	0.135
Lang	17	0.315
MU	0	*****
Urb	23	0.060

A minor digression on census procedures. “Conventional enumeration” means that respondents were asked to fill out the forms and hold them for collection by an enumerator; this process was used in largely rural areas, particularly in the west. Conv is the percentage of persons living in areas that were conventionally enumerated. (In urban areas, forms were to be mailed back.) The undercount in 1980 was relatively high in rural areas, probably due to incomplete maps and address lists; that may be why conv is such a powerful explanatory variable.

We did an additional simulation with PEP 2-9 taken as truth, allowing percent urban to be selected as an explanatory variable. The results are shown in Table 10. Again, the percent conventionally enumerated comes in as does the percent minority. Otherwise, there is a fair degree of inconsistency. And the much-maligned percent urban is chosen more often than 5 of EKT’s variables, including the central-city indicator. The data do not determine the model.

7.4 The Regression Model at Trial

As statisticians, we are intrigued by arguments about regression. However, the court was not impressed:

“In their rebuttal case, the plaintiffs argued that the application of regression analysis to the undercount estimates derived from the PEP would enable the Bureau to use the PEP to accurately adjust the 1980 census. However, both plaintiffs’ and defendants’ experts agreed that regression analysis will not in any way alleviate the bias in the PEP and plaintiffs apparently do not contend otherwise. In short, while regression analysis may remove some of the random sampling error in the PEP, regression analysis will not reduce the substantial errors in the PEP caused by erroneous matches, the untested assumptions made with respect to the unresolved cases, and correlation bias. Moreover, the overwhelming weight of the evidence supports the conclusions of defendants’ experts that the principal difficulties with the PEP stem from these biases rather than from sampling error.” (674 F Supp 1103, footnotes and citations omitted.)

8. SUMMARY AND CONCLUSION

Ericksen, Kadane, and Tukey argue that they can improve on the 1980 census counts by statistical adjustment. They seem now to agree that adjustments would not have been justified for subareas of the 66 PEP study areas. With respect to the 66 areas themselves, disagreement remains. In our opinion, success of any of EKT's proposed adjustments rides on unverified and implausible assumptions—about missing data, undercount mechanisms, bias in PEP, and stochastic errors in regression models. Changing the assumptions changes the results, and taking averages over various sets of assumptions does not, at least in our opinion, make the problem go away. EKT conclude (p. 943).

“We believe that the Census Bureau creates political difficulties for itself when it ignores the undercount. The bureau will put itself in a better position by making its best effort, using available statistical and demographic methods, to adjust for the undercount. Errors will remain, but they will be smaller and we will no longer know in advance who is losing money and power because of the undercount.”

This political analysis has merit, but there are caveats. We think it quite unfair to say that the Bureau has ignored the undercount. Nor are the Bureau's political difficulties entirely of its own creation. Adjustments can indeed be devised to satisfy particular groups or settle individual law suits. However, the census is used to share out fixed resources, so there will always be losers as well as winners. These will have little trouble identifying themselves, after the fact if not before. And up to now, the goal of improving on the accuracy of the census by statistical adjustment has proved illusory.

9. HOW DID THE COURT RULE?

At the time of writing, litigation about the 1990 census goes on. With respect to the 1980 census, however, the court ruled for the defendants on all the issues. We quote from the digest and opinion *Cuomo et al. v. Baldrige et al.* 674 F. Supp. 1089-1108 (SDNY 1987).

“State, city, and their officials brought action against Secretary of Commerce, Director of the Bureau of the Census, and other officials seeking statistical adjustment of 1980 decennial census. The District Court, Sprizzo, J., held that state and city failed to establish that statistical adjustment of decennial census was technically feasible.”

“... it is essential to any such adjustment that a technically feasible adjustment methodology exist which gives a truer picture of the United States population on a state-by-state basis for apportionment purposes, and a sub-state-by-sub-state basis for federal funding purposes ... If it does not, then no adjustment can or should be made ... because ... both congressional seats and revenue sharing funds are fixed quantities, and an increase in the population in one state or sub-state area will adversely affect the shares of other localities ...

“Notwithstanding the complexity of the facts ... this action presents one issue to be resolved by the Court: whether the plaintiffs have sustained their burden of proving that a statistical adjustment of the 1980 census will result in a more accurate picture of the proportional distribution of the population of the United States on state-by-state and sub-state-by-sub-state basis than the unadjusted census. The Court finds as a matter of fact that the plaintiffs have not sustained that burden, and the action must therefore be dismissed ...”

ACKNOWLEDGEMENTS

Freedman is working for the Department of Justice on matters arising from the 1990 census. However, the views expressed in this article may not be shared by the Department. Helpful comments were made by L. Bazel (San Francisco), P. Diaconis (Harvard), S. Klein (RAND) and A. Tversky (Stanford). Research partially supported by NSF Grant DMS 86-01634.

APPENDIX

Synthetic Estimation and Loss Functions

Synthetic Estimation

Section 5 in Wolter and Causey (1991) describes their empirical proof that synthetic adjustment would have brought the 1980 census closer to truth. The evidence is a simulation study: the “census” and “truth” are both defined in terms of an artificial reference population developed by Isaki *et al.* (1987). However, the argument depends rather strongly on the reference population, as shown by Passel (1987). The object here is to sketch a variation on one of Passel’s examples. Indeed, if the reference population is defined by using PEP 2-9 to correct the 1980 census, then synthetic adjustment moves the counts farther from truth.

Table 11 shows the data for the four census regions – Northeast, Midwest, South, and West. With squared differences in population shares weighted by size,

r.m.s. difference between Synthetic B and PEP 2-9 = 0.21 of 1%.

(13)

r.m.s. difference between the Census and PEP 2-9 = 0.15 of 1%.

(14)

PEP 2-9 is rather close to the “average preferred PEP” in Table 2. In that table, the census was closer to synthetic B than to the PEP estimates. In Table 11, the census is closer to PEP, and synthetic B is the outlier. The difference between the two tables seems to be the disaggregation. Table 2 disaggregates the U.S. by race and ethnicity; Table 11, according to conventional census geography.

Of course, using another disaggregation or a different synthetic adjustment could reverse the comparisons yet again; so could a change in the loss function. To illustrate the possibilities, consider adjusting the 66 PEP study areas, rather than four regions. Keep PEP 2-9 as ‘truth.’ Using the loss function (17), the census is preferred to synthetic B, by a little. Using (16), synthetic B shows a much smaller loss than the census.

Table 11
Population Shares from The Census, The Synthetic B Estimates,
and PEP 2-9, in Percent; Census Counts, in 1,000s

	Northeast	Midwest	South	West	Total
Synthetic B – PEP 2-9	.08%	.03%	.24%	– .35%	.00%
Census – PEP 2-9	.10%	.06%	.12%	– .28%	.00%
PEP 2-9	21.59%	25.92%	33.15%	19.34%	100.00%
Synthetic B	21.67%	25.95%	33.39%	18.99%	100.00%
Census	21.69%	25.98%	33.27%	19.06%	100.00%
Census count	49,135	58,866	75,372	43,172	226,545

Loss Functions

Proponents of adjusting the 1990 census make analytic arguments based on loss functions: see Wolter and Causey (1991) or Ericksen, Estrada, Tukey and Wolter (1991, p. 20 of the main report; Appendices G and H). The essence of argument can be summarized in the lemma which follows. To set up the notation: the country is divided into n areas, indexed by i ; c_i is the census count in area i and t_i is the true count. The “synthetic estimate” for area i is $x_i = \lambda c_i$, where the “adjustment factor” λ is computed from other data.

Lemma. For $i = 1, \dots, n$ let $c_i > 0$ and $t_i > 0$. Let $0 < \lambda < \infty$ and $x_i = \lambda c_i$. (15)

Then

$$\sum_{i=1}^n (x_i - t_i)^2 / c_i \quad (16)$$

is minimized when

$$\lambda = \left[\sum_{i=1}^n t_i \right] / \left[\sum_{i=1}^n c_i \right].$$

The proof is omitted as trivial. The “loss function” defined by (16) differs in detail from the one used in (7), (8), (9), (13) and (14), which can be written as

$$\sum_{i=1}^n \frac{c_i}{C} \left[\frac{x_i}{X} - \frac{t_i}{T} \right]^2, \quad (17)$$

with

$$C = \sum_{i=1}^n c_i, \quad X = \sum_{i=1}^n x_i, \quad T = \sum_{i=1}^n t_i.$$

The loss function (17) emphasizes shares while (16) emphasizes counts; furthermore, (17) puts more weight on large sub-populations while (16) does the opposite, due to the division by c_i . We are not particularly attached to (17), and see no good way to choose one loss function rather than another.

Lemma (15) is mathematically correct, but it is so far removed from the realities of adjusting the 1990 census that it seems virtually irrelevant. In this connection, there are four points to consider:

- (a) The true population total T is unknown; Wolter and Causey attempt to deal with this problem, but the example in Table 11 refutes their argument: synthetic adjustment makes the 1980 census less accurate.
- (b) Synthetic estimates do not perform well under aggregation.
- (c) At the block level, rounding error may dominate.
- (d) Loss functions only capture part of the policy problem, and may obscure more than they reveal.

Points (b), (c) and (d) will be discussed in more detail; but first, a brief review of proposed methods for adjusting the 1990 census. The population is divided into 1,392 “post strata,” e.g. male hispanic renters age 30-44 in central cities in the Pacific Division. Index these post strata by $j = 1, \dots, 1,392$. For each post stratum j , an adjustment factor λ_j is computed by capture-recapture techniques from data collected in a Post Enumeration Survey (Freedman 1991).

The 1,392 factors are used to adjust all small-area counts as follows. Fix an area, e.g. a town. This area will intersect many of the post strata. The census count for each area \times post stratum intersection is multiplied by the corresponding λ_j , and the products are summed. In other words, subpopulations are adjusted by the synthetic method, and synthetic estimates are aggregated to obtain totals for small areas.

This completes a sketch of the adjustment process, and we return to points (b), (c) and (d).

(b) Synthetic estimates do not perform well under aggregation. This was already pointed out by Fellegi (1985). See Cohen and Citro (1985, p. 318). For another example, see Tables 3 to 5 above.

(c) At the block level, rounding error may dominate. Census adjustment would in fact be done at the block level. (A “block” is the smallest unit of census geography; there are 6.5 million blocks in the country.) A typical block in an urban area may intersect 25 post strata; each block \times post stratum intersection contains only a handful of people. Multiplying by an adjustment factor means adding or subtracting a fractional number of people, and the fractions would be rounded. The next example illustrates how rounding error may offset any advantage from synthetic adjustment.

Suppose there are n “areas” to adjust; these could be viewed as blocks intersected with one fixed post stratum. Suppose each of these areas has the same census count, c . Fix $m < n$. Suppose that in each of m areas, the census has missed one person; in the remaining $n - m$ areas, the census count is exactly right. In all, there is an undercount of m people. These facts are considered as known; but it is not known which blocks have the missing people. According to (16),

$$\text{loss from using the unadjusted census} = m/c. \quad (18)$$

Adjustment would proceed as follows: choose m areas at random, and add one person to each of these areas. Clearly, the expected loss from adjusting is

$$\begin{aligned} \frac{m}{n} \cdot m \cdot 0 + \left(1 - \frac{m}{n}\right) \cdot m \cdot \frac{1}{c} + \frac{m}{n} \cdot (n - m) \cdot \frac{1}{c} + \left(1 - \frac{m}{n}\right) \cdot (n - m) \cdot 0 \\ = 2 \left(1 - \frac{m}{n}\right) \cdot \frac{m}{c}. \end{aligned} \quad (19)$$

Lemma. If $m < n/2$, there is an expected net loss from synthetic adjustment.

Proof. If $m < n/2$, then

$$2 \left(1 - \frac{m}{n}\right) \cdot \frac{m}{c} > \frac{m}{c}. \quad (20)$$

Of course, this example is almost as stylized as Lemma (15). In short, the value of census adjustment cannot be established by *a priori* argument.

(d) Loss functions capture only part of the policy problem, and may obscure more than they reveal. To begin with an example, suppose that the census is in error, and the main impact of that error is to transfer a congressional seat from California to Pennsylvania. There is a gain for Pennsylvania, and a loss for California. There may be a net social loss from this misallocation, but attempting to quantify that loss by (16) – or any similar formula – seems quite simplistic.

We now present another example to illustrate point (d). To focus the issue, suppose the census undercount is largely confined to blacks and hispanics in New York, Chicago, Houston and Los Angeles. The census, by assumption, under-estimates the share of the population living in these four cities, and adjustment will partly correct that error.

Due to its reliance on the synthetic method, however, adjustment will change population shares everywhere. Areas which are heavily black and hispanic will have their population shares artificially increased, at the expense of other areas. This will be so even in regions of the country where the census was accurate.

In this example, the distribution of resources between the four cities and other areas may be made fairer by adjustment – at the expense of distortions introduced everywhere else. The loss-function approach slides over this difficulty. Balancing inequities is a political problem, not easily resolved by a statistical formula.

Some observers may consider the example to be extreme. However, the Post Enumeration Survey only samples 5,000 blocks, and there are 39,000 jurisdictions to adjust. Real information about the undercount is necessarily confined to relatively few localities. Adjustments for other areas must therefore be based largely on theory rather than data.

REFERENCES

- CITRO, C.F., and COHEN M.L. (Eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C. National Academy Press.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 80, 98-131.
- ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927-943.
- ERICKSEN, E.P., ESTRADA, L.F., TUKEY, J.W., and WOLTER, K.M. (1991). Report on the 1990 Decennial Census and the Post Enumeration Survey, submitted to the Secretary of the Department of Commerce, June 22, 1991.
- FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). *The Coverage of the Population in the 1980 Census*. Washington, D.C.: U.S. Department of Commerce, Government Printing Office.
- FELLEGI, I. (1985). Comment. *Journal of the American Statistical Association*, 80, 116-119.
- FREEDMAN, D.A. (1991). Adjusting the 1990 census. *Science*, 252, 1233-1236. Copyright 1991 by the AAAS. Excerpted by permission.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 1-39.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

- ISAKI, C., DIFFENDAH, G., and SCHULTZ, L. (1987). Report on statistical synthetic estimation for small areas. Technical report, Bureau of the Census.
- PASSEL, J. (1987). A note about synthetic estimates of undercount. Memorandum, U.S. Bureau of the Census.
- SCHIRM, A.L., and PRESTON, J. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.
- SCHIRM, A.L. (1991). The effects of census undercount adjustment on congressional apportionment. *Journal of the American Statistical Association*, 86, 526-541.
- WOLTER, K. (1986). Comment. *Statistical Science*, 1, 24-28.
- WOLTER, K. (1991). Accounting for America's uncouned and miscounted. *Science*, 253, 12-15.
- WOLTER, K., and CAUSEY, B. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.
- YLVISAKER, D. (1991). A look back at TARO. Technical report, Department of Mathematics, UCLA.

COMMENT

STEPHEN E. FIENBERG¹

Freedman and Navidi give their current thought-provoking retrospective on the issue of undercount in the 1980 U.S. decennial census. Unfortunately they fail to address the question posed in the title of their paper and instead attempt to vindicate their views expressed earlier in Freedman and Navidi (1986) and to rebut commentaries on these views by others. Their theme is a familiar one to those who have read earlier versions of the debate connected with the "1980 lawsuit" over adjustment: The census is very complex and only a small undercount is thought to remain; adjustment utilizes statistical modelling that relies on unverifiable assumptions; a bad adjustment may be worse than nothing.

I disagree with many of the views expressed by the authors and believe that they distort both what should have been at issue with respect to 1980 and what appears to be at issue in litigation currently pending over correction of the 1990 census. In the following, I attempt to explain my differences with the authors and give my perspective on two questions: the one raised in the title and the one implicit in the material introduced regarding the 1990 census. (Note: The author played no part in the litigation over the adjustment of the 1980 census but he is working with the City of New York and other plaintiffs in litigation stemming from the decision by the Department of Commerce not to adjust the results of the 1990 census.)

1. The Title and the Paper Address Two Different Issues

Should we have adjusted the census of 1980? The only sensible way to answer this question in my mind is to ask it in the context of the evidence available at the time, or at least available when the issue was being adjudicated by the courts. As such, the description of the issues identified by the Bureau of Census and presented in the opening section of the paper are important, although they had little to do with the original decision not to adjust in 1980 made by the Director of the Bureau in advance of the availability of coverage information.

The remainder of the paper, however, does not deal with this question. Rather, it addresses the continued attempt by advocates for the two sides to marshal evidence to support their positions from the litigation. In essence, the authors are asking a question about the current evidence in support of a decade old decision. As with all statistical issues, continued data analysis and retrospection can update our judgment on the answer to such a question and thus the authors' effort to revisit the evidence connected with the 1980 census yet again is to be applauded.

We can thus turn to the framing of the question to be answered. For me, the judge's statement of the issues at trial falls short of the mark, as does Freedman and Navidi's description of the undercount issue. They imply that the only real issue is the accuracy of the adjustment process and that there is only a potentially small undercount about which we should be worried. Neither could be further from the truth. At issue is both the accuracy of the census and the adjustment process. And, it is the substantial *differential* undercount, *i.e.* the difference between the undercount for Blacks and the undercount for non-Blacks and between Hispanic and non-Hispanic, that is important when we come to assess census accuracy. This is because census figures are typically used to divide resources among groups in the population, resources such as seats in the U.S. House of Representatives; seats in state legislatures; federal funds; and so on.

¹ Stephen E. Fienberg, York University, North York, Ontario, Canada M3J 1P3.

Using the method of demographic analysis the Census Bureau has documented that, from 1940 through 1980, the difference in the rate of undercount for Blacks and non-Blacks has remained roughly constant, somewhere between 5% and 6% even though the overall undercount declined from 5.6% to 1.4% (see Fay *et al.*, 1988). The 1.4% figure does not mean that the census correctly counted over 98% of the U.S. population in 1980. Rather 1.4% represents the *net* undercount, which can be thought of as the difference between the actual undercount (consisting of missed individuals or omissions) and the overcount (erroneous enumerations and duplications). Even if the errors of overcount and undercount balanced perfectly at the national level, thus producing a 0% national undercount, we might still have a differential undercount problem. For the 1980 census, the Bureau determined that there were 6 million erroneous enumerations in the census, of which as many as 1 million were fabrications, and as many as 2.5 million people were erroneously included twice at the same location. Given the Bureau's report of a net undercount of 1.4% or 3.2 million people in 1980, we have an estimate of 9.2 million omissions (people who were missed) from the 1980 census count. By adding omissions to erroneous enumerations we get a total of 15.2 million errors in counting individuals, which corresponds to almost 7% of the official 1980 census total. To me, this level of error in the census represents a major problem that must be addressed when we talk about the appropriateness of adjustment in 1980. Of course, shortly after the 1980 census was completed the Census Bureau painted a much rosier picture of the accuracy of the raw census counts. Perhaps, in keeping with the literal meaning of the title of this paper, Freedman and Navidi wish us to accept as accurate what we now know to have been a seriously incomplete assessment on the part of the Census Bureau. I hope this is not the case. We now know much more about the level of the error in the raw census counts from 1980. The residual issue is whether we have any better information about the various forms of adjusted counts given the passage of a decade.

2. Facts and Theorems

The present paper is full of statements about the accuracy of the census adjustment procedures. When it comes to stating and proving theorems, I have no doubt that Freedman and Navidi will get them correct. The relevance of such theorems for census adjustment is a different issue.

Freedman and Navidi present a simple and seemingly compelling counterexample to the Schirm-Preston theorem on synthetic adjustment. It is certainly true that the overall totals for state A and B in their example are correct in the census and incorrect in the synthetic adjustment, although barely so. But it is also true that the large shift of the counts of Whites and Blacks in state B is what I understand that an adjustment is designed to accomplish and it does so at the expense of a minor perturbation in State A. Moreover, if the fictional state B is like those in the real U.S., the distributive accuracy of the synthetic data for geographic areas within State B is much improved while that within State A seems not to be seriously affected. Freedman and Navidi also offer their conclusion in the form of a parable to which I respond with one of my own. Small overall undercounts can hide a multiplicity of censal errors, ones that tend to "balance" in the aggregate but exact a heavy toll from states with large hard-to-count minority populations.

I also found the evidence from the Schirm-Preston simulations far more credible than did Freedman and Navidi and wonder whether this may be related to the corrected version of the Schirm-Preston theorem that is referred to as holding under more complicated sets of conditions involving weighted averages. What I am asking is whether the corrected theorem is more relevant to the real problems of undercount in the U.S. than Freedman and Navidi's counterexample.

3. Issues in Dispute with EKT

Freedman and Navidi spend much time rehashing the issue of the multiplicity of PEP series and by stressing the variations amongst them. While there is some merit in the position that there is not a clear and overwhelming choice from amongst the adjustment alternatives, it may still be the case that several choices would be superior to an unadjusted census. The authors focus on the variation amongst the full set of 12 alternatives, some of which to me are implausible given the assumptions that they rely upon. Even though I do find the arguments in support of the use of synthetic adjustments reasonable, I do agree with the authors that there is a clear difference between the synthetic and PEP adjustments.

Where are we left in this debate? I find the conclusion of Ericksen, Kadane, and Tukey compelling even though I agree with Freedman and Navidi that issues remain about the specific choice of techniques favored by EKT. Freedman and Navidi argue that their principal claim is irrelevant to the issue of accuracy. I disagree. Perhaps the authors believe that the millions of uncounted people that virtually all agree were missed in 1980 are still out hiding in the foothills of South Dakota, or in some other state with few minorities.

A familiar theme in various writings by one of the present authors is the problems that arise when assumptions are not satisfied. Here again the authors pursue this theme with respect to the linear equation used for smoothing. They appear to argue that either all assumptions must be perfectly justified or "all bets are off". Nothing could be further from the truth. Surely they don't expect anyone to believe the argument that the eruption of Mt. St. Helens interfered with census taking in a serious way and thereby undercuts the usefulness of the smoothing approach. Similarly, their notion that precise specification of predictor variables is crucial to the accuracy of smoothing is also something with which I take issue. Finally, I read the report by Ylvisaker (1991) who reexamined data from the trial census in Los Angeles in preparation for 1990, but I could not find the evidence Freedman and Navidi state is supportive of their claim that smoothing increases variability.

I do believe with Freedman and Navidi that the census process is enormously complex and that the approach to adjustment that was proposed in connection with the litigation over the 1980 census is far from flawless. Yet I still find their arguments exaggerated and they tend to obscure the old maxim that "the best is the enemy of the good." Of course the assumptions are not satisfied. Of course one could produce a better way to adjust that does not suffer from all of the flaws in the methods advocated by EKT. But this does not mean that adjustment with these flawed methods would not have been an improvement over the badly flawed unadjusted counts.

4. Adjustment in 1990

At various points throughout the paper the authors allude to comparable issues and imponderables in connection with adjustment in the 1990 census. I think that the reader should make a clear distinction between the methods used in connection with analyses presented as part of the 1980 lawsuit and those used as an integral part of the 1990 census. Many of the problems encountered by those who attempted to prepare adjusted figures in 1980 have clearly been overcome and the debate over adjustment in 1990 has become much sharper in its focus. Moreover, unlike in 1980, the key statistical methodologists at the Census Bureau, and the Director herself, found the adjustment methods used in 1990 justifiable and they recommended proceeding with an adjusted census. The statisticians were overruled by the Secretary of Commerce. The matter is now in the hands of the court once again.

Freedman and Navidi do not state their position regarding the use of adjustment techniques for the 1990 census, but Freedman (1991) makes quite clear that his judgment from 1980 has not changed. I disagree with this view. There may well be reason to argue, as the authors do, that the Census Bureau should not have adjusted the census in 1980. But 1990 is another matter. In June, the General Accounting Office, an investigative arm of the U.S. Congress, reported that there were 25.4 million gross errors in 1990 census, or about 10.4% of the resident population. The Bureau estimates that the net undercount was about 5 million people and that the differential undercount was the largest since the Bureau began to estimate it beginning with the 1940 census. Methodology for carrying out an adjustment in 1990 is much improved relative to that at issue in 1980. In my view, the results of the Census Bureau's evaluation studies clearly supported the use of adjustment for the 1990 census results. Perhaps the judge this time will see the issue of adjustment differently than the the way that Freedman and Navidi tend to frame it.

COMMENT

IVAN P. FELLEGI¹

Freedman and Navidi provide a very thorough and lucid description of the considerations and arguments surrounding the adjustment debate for the 1980 US Census. These arguments focus on the quality of population counts and population distributions for Census day 1980. Furthermore, it is taken as given that whatever decision is made on adjustment, based on this consideration of population counts, will be applied to the complete census database and therefore to all the outputs flowing from it. Rather than commenting in detail on the arguments of the protagonists in this debate (though I am of the view that the correct decision was made for the 1980 Census), I would like to offer, from a Canadian viewpoint, some thoughts that suggest a broader frame of reference for the adjustment debate.

1. The Census is Much More Than a Head Count

Ever since the age of modern census taking began, the objective has always been more than the provision of an accurate count of the population. Yet the increasingly impressive literature dealing with the issue of adjusting the census tries to assess the relative advantages of alternative courses of action solely from the point of view of estimating the total number (and proportion) of persons living in a set of areas. I understand, of course, why this is so: (a) the problems involved are difficult enough as it is, and (b) so much money and political power is associated with the population counts (or estimates).

I will come back to point (b) above. As far as (a) is concerned, I think it would not be a scientifically defensible position to adjust the census by whatever method without taking into account the impact of such an action on the multitude of uses of census data. Indeed, I believe that if the objective of the census was restricted to estimating population totals and distributions, we would most likely (at least in Canada) try to find quite different methodologies to fulfil such a very different role. Given the multiplicity of objectives served by the Census, the fact that this multivariate and rich data base is difficult to model is not an adequate excuse for dealing with the much simpler issue of population counts and then uncritically applying the conclusions to the entire data base.

2. Point-in-time Precision of Population Counts May not Be the Relevant Measure for the Intercensal Distribution of Federal Funds and Power

There seems to be a preoccupation with exquisite precision of population counts and distributions **in the census year**. Of course a periodic stock-taking, providing good and comparable data for small areas and/or small population groups is a main justification for the expense involved in taking a census. But the excessive (it seems to me) preoccupation with the precision of the census count is motivated by equity considerations: a great deal of money and political power is distributed based, in part, on population numbers. Let us examine these two equity issues in turn.

First, dealing with the distribution of funds, indeed substantial sums are distributed in Canada from the federal government to provinces based on formulae that are very sensitive to population numbers and distributions. However, two points are of great significance from the point of view of census adjustments.

¹ Ivan P. Fellegi, Chief Statistician of Canada, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

- (1) The formulae use a large array of statistical information (most of it derived from sources other than the census), only one component of which is population. It is well known that several of the other components are subject to significant sampling and non-sampling errors. It is an open question whether any reasonable loss function designed to assess the combined impact of all the errors involved would be materially improved even if the census errors could be entirely eliminated.
- (2) Even more important, if the adjusted population numbers result in a smaller loss function, or if more generally they are assessed to be closer to the truth for a significant majority of the areas involved, then these can serve as the basis for improved **population estimates** (and not just in census years) without adjusting the entire multivariate census data base. In Canada (and in the United States) there is a long history of publishing estimates of the census undercount. Serious consideration is, indeed, being given in Canada to taking the next step: publish the census results as taken, and have a set of official population estimates which takes account of the known census undercount. After all, in non-census years the official population estimates incorporate a wide range of estimation techniques – some of them having errors at least as large as the likely errors of undercount estimates (even model-based ones). It may be scientifically quite appropriate to publish the best available population estimates in both census and non-census years – whether or not these estimates coincide with the census counts in census years. It may well be that legislation, or regulations under existing legislation, have to be amended to permit the use, particularly in a census year, of population estimates different from those directly derived from the census. But (a) that has little to do with the scientific arguments involved in the adjustment debate, and (b) it is more honest than relabelling the “adjusted” census counts to be “the” census counts simply because the law might require the latter.

The arguments are different in respect of the distribution of political power based on census counts, although the fixation on point-in-time precision seems to me to be equally misplaced. Indeed, the census population figures are also used to distribute seats in the House of Commons in Canada (and in the House of Representatives in the USA). However, the distribution of seats based on the census is used for ten years. During those years typically massive population shifts occur. Leaving aside the interpretation of laws, it seems to me that the substantive question is whether a suitably defined loss function, designed to capture the average deviation from the objective of “one person one vote” **over a ten year period**, would be materially reduced if the census counts were adjusted for the estimated undercount. I have not made such a calculation. However, it seems to me that the range of population shifts over ten years are substantially larger than the range of estimated undercounts. I would therefore, speculate that even apparently significant potential census year adjustments (and corresponding shifts in the allocation of seats in the legislature) are relatively less significant than the deviations from the “one person one vote” rule occasioned by migration over a ten year period. Since this particular use of the census is mandated by the constitution, changing the law is not an option. But a scientifically informed debate regarding the appropriate interpretation of the constitution is very much in order – taking full account of the two main causes of deviation from equity in political representation during the ten year intercensal period: census errors and population shifts (mostly migration).

3. Conclusion

- (a) The census is a multivariate integrated data base. The case for “adjusting” it is far from obvious, even if (a big if) the simplest variable involved – the count – can be improved by doing so.
- (b) If a set of population estimates that are judged to be better than the census counts (according to suitably defined criteria) can indeed be generated, these should be produced and used, without necessarily adjusting the entire census data base. The criteria should relate to the set of areas (and other breakdowns) for which estimates are required.
- (c) If the law requires “census” derived population counts when in fact substantively the best available population estimates are called for, it would appear to be preferable to try to change the law rather than to adjust (in effect weight) the entire census data base to agree with estimated population numbers – solely in order to be able to refer to the population estimates as “the census”.
- (d) Equity considerations, both in terms of the distribution of federal funds and political representation, apply to the entire intercensal period, not simply for the year of the census. They should be studied using models that take full account of this fact.

COMMENT

N. CRESSIE¹

A critical assessment of our past successes and failures makes us better equipped to provide future successes. Missing data and matching problems in the 1980 Post Enumeration Program were major impediments to a successful adjustment of the 1980 U.S. Decennial Census. A court case, *Cuomo et al. versus Baldridge*, was brought by New York State and others to require the Census Bureau to adjust the 1980 Census numbers for undercount. Testimony from Barbara Bailar, then Associate Director of Statistical Standards and Methodology at the Census Bureau, and Kirk Wolter, then Chief of the Statistical Research Division at the Census Bureau, made it clear that 1980 data and methods were inadequate for an accurate adjustment of the whole country.

In 1987, Judge Sprizzo ruled against New York. However, that decision did not make the differential undercount go away; even the judge in his ruling acknowledged its presence. There is little disagreement that, differentially by race, national U.S. Census numbers have been persistently too small. Using demographic methods, the following estimates are available.

- 1950: Black (and other non whites) demographically estimated undercount was 9.7%. White demographically estimated undercount was 2.5%. (Siegel 1974, Table 3).
- 1960: Black (only) demographically estimated undercount was 8.0%. White (and other races) demographically estimated undercount was 2.1%. (Siegel 1974, Table 2, set D estimates).
- 1970: Black (only) demographically estimated undercount was 7.6%. White (and other races) demographically estimated undercount was 1.5%. (Passel, Siegel and Robinson 1982, Table 1).
- 1980: Black (only) demographically estimated undercount was 5.3%. White (and other races) demographically estimated undercount was -0.2%. (Passel and Robinson 1984, Table 2).

Further, there is little disagreement that racial composition is different within administrative regions (both large and small) across the U.S.A. The consequence of these two virtually undeniable facts is that undercount will be differential across administrative regions, leading to an unrepresentative geographic/racial profile of the nation and an unfair apportioning of political and financial resources. So, Freedman and Navidi state in their introduction “... If the undercount can be estimated with good accuracy, especially at the local level, adjustments can – and should – be made to improve the census.”

Almost everyone agrees there is a problem. The adage, “If it ain’t broke don’t fix it,” does not apply here. It is an uncomfortable defence for a statistics professional to argue that uncontrolled-for biases and errors will not allow an adjustment for an undercount that is known to be there and known to be damaging. During the early 1980s, Bailar and Wolter established the Undercount Research Staff within the Statistical Research Division of the Census Bureau. Staff members have produced high-quality research that demonstrated “that it is technically feasible to correct the 1990 Census for differential undercoverage”: (Childers *et al.* 1987).

It is time for Freedman and Navidi to relinquish their role as devil’s advocates; it is time for them to put their knowledge and talents into a constructive mode; and it is time for them to say what they mean by “good accuracy,” “local level,” and various other qualitative affirmations. The adversarial atmosphere of the courts has spilled over into the various articles,

¹ N. Cressie, Department of Statistics, Iowa State University, Ames, IA U.S.A. 50011.

comments, and rejoinders we have seen on census undercount in the last 10 years. To solve a problem as hard as adjustment for undercount, the common goal needs to be recognized. From there, debate should center around differences on how that goal might be reached. If Freedman and Navidi's position is that the goal is impossible to reach (which is what they seem to have implied over the years), then it should be stated.

For the rest of this comment, I shall address a number of important technical matters that were raised by Freedman and Navidi (1986) and now, surprisingly, again in the article under discussion. In 1990, I presented a paper at the Census Bureau's Annual Research Conference (Cressie 1990) that David Freedman was invited to discuss. At the last minute, he was unable to attend the conference but I continued to send him the discussion version and the final version and invited his comments. The paper is rather technical but addresses, successfully I believe, several major criticisms made by Freedman and Navidi (1986) of the statistical modeling approach to undercount adjustment.

First, the paper expresses a preference for the "stratification approach" over the "regression approach". Stratification is a special case of regression where the explanatory variables are restricted to 1 and 0, indicating presence or absence in a particular (demographic) stratum. There is little disagreement that undercount is differential across sex \times age \times race/ethnicity strata. Because the Census Bureau was committed to a regression approach, the bulk of the paper addressed the more general problem.

Second, if one allows the regression error (see Freedman and Navidi's *e.g.* (2)) to be dependent, such models can absorb bias and misspecification into the error term. The important concept to maintain is that true undercount in regions is unknown and the ignorance is quantified into a probability model. The goal is *not* estimation of the coefficients β but *prediction* of the undercount. With an error term that does not have to be independent and identically distributed, this prediction is insensitive to misspecification (see also Cressie 1991, Chapter 3).

Third, the inconsistency of the model to changes in geographic level is addressed by modeling adjustment factors, not undercounts, and by assuming the variance of the regression error of a particular area is inversely proportional to that area's population. This assumption is justified, from both a Bayesian and frequentist point of view, in Cressie (1989).

Fourth, the effect of estimation of variance-covariance parameters can be taken into account by modifying the results of Prasad and Rao (1990) to a multivariate context. One could also use a parametric bootstrap, by generating data from the estimated model, re-estimating all parameters, and repredicting the undercount.

Finally, it is acknowledged that all preceding model-based methods will likely do poorly if the model does not fit. Diagnostic methods are crucial to the success of statistical model-based adjustments for undercount.

There is room for critical assessment of our past successes and failures. It is time to move on and solve this monumentally important problem with cutting-edge technology. A well designed, well implemented, and quality-assured 1990 Post Enumeration Survey with excellent computer matching and precise geography make the 1980 case look very different indeed. It is my opinion that adjustment can now be successfully carried out at the state level. Research and debate on whether that success can be carried down to lower levels of geography deserves our collective resources (*e.g.* Tukey 1983; Cressie 1988; Wolter and Causey 1991). *Expected* losses (or risks) can be used to measure the efficacy of adjustment procedures. Cressie (1988) gives sufficient conditions under which synthetic adjustment improves over census count; those conditions were satisfied in the 1980 Census and PEP 3-8 series.

ADDITIONAL REFERENCES

- CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., and WOLTER, K. (1987). Paper presented in the Session *1990 Census Undercount and Adjustment*, American Statistical Association Annual Meetings, San Francisco, California.
- CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 205-222.
- CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- CRESSIE, N. (1990). Weighted smoothing of estimated undercount. *Proceedings of Bureau of the Census 1990 Annual Research Conference*. Bureau of the Census, Washington, D.C., 301-325.
- CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.
- PASSELL, J.S., and ROBINSON, J.G. (1984). Revised estimates of the coverage of the population in the 1980 Census based on demographic analysis: A report on work in progress, in *Proceedings of the Social Statistics Section, American Statistical Association*, 160-165.
- PASSELL, J.S., SIEGEL, J.S., and ROBINSON, J.G. (1982). *Coverage of the National Population in the 1980 Census, by Age, Sex and Race: Preliminary Estimates by Demographic Analysis*. Current Population Reports, Special Studies P-23, No. 115. Bureau of the Census, Washington, D.C.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.
- SIEGEL, J.S. (1974). Estimates of the coverage of the population by sex, race and age in the 1970 Census. *Demography*, 11, 1-23.
- TUKEY, J.W. (1983). Affidavit, presented to District Court, Southern district of New York. *Cuomo et al. versus Baldridge*. 80 Civ. 4550 (JES).

COMMENT

ALLEN L. SCHIRM and SAMUEL H. PRESTON¹

1. Introduction

We thank the editor for inviting us to comment on this provocative article by Freedman and Navidi (hereafter, "F and N") and continue this important policy debate. Our comment mainly responds to F and N's criticisms of our earlier research (Schirm and Preston 1987; hereafter, S and P). Although we disagree with much of F and N's critique of Ericksen, Kadane and Tukey (1989), we leave to the authors of that article the task of defending their work.

We disagree with many of F and N's specific criticisms of S and P. Before discussing our detailed responses, we want to take a broader perspective and view our article and F and N's criticisms of it in their entirety.

F and N wrongly characterize our article in stating that we "present two major arguments, one analytical and one based on simulation." In fact, we presented three analytical results. F and N criticize only one, and a minor one at that. Our most important analytical result suggests that synthetic adjustment would likely have improved the accuracy of the population distribution in 1980. As for our simulations, they were not intended to support any one argument. Instead, we simulated an extremely wide variety of circumstances to permit us to address several questions about synthetic adjustment and its effects. We found, however, that adjustment would have improved accuracy under all conditions simulated, including highly unfavorable circumstances.

2. Analytical Results

In S and P, we presented three analytical results. All three are mathematically correct. However, the second result – the sole target of F and N's criticism – is, as we stated in our article, potentially "misleading because it ignores influences on overall adjustment success of systematic relationships between variations across states in census coverage for a group and differences between groups in how they are distributed across states." Our third result, which is clearly the focus of our algebraic analysis and which does not depend on the second result, addresses this issue and takes into account the patterns of variations in undercounts across states. Although potentially misleading, we presented the second result to illustrate more forcefully a key implication of our third result, that systematic variations in state undercounts can matter.

Our second analytical result suggests that the effect of adjustment for a given state hinges on how "close" the state's undercounts are to the national undercounts. Contrary to F and N's claim, our second analytical result is mathematically correct. F and N are able to dispute our finding only because they choose to define "close" without regard to our precise definition. Thus, F and N's "counterexample" to our second result does not pertain to that result at all, since their example violates the conditions that we derived and stated precisely in the appendix to our article. To repeat that result, we showed that the estimated proportion of the total national population residing in state i is made more accurate by adjustment if

$$\left| \sum_{j=1}^J \frac{N_{ji}^C}{N_{..}^C} \left(\frac{N_{..}^T}{N_{..}^C} - \frac{N_{ji}^T}{N_{ji}^C} \right) \right| > \left| \sum_{j=1}^J \frac{N_{ji}^C}{N_{..}^C} \left(\frac{N_{j.}^T}{N_{j.}^C} - \frac{N_{ji}^T}{N_{ji}^C} \right) \right|,$$

¹ Allen L. Schirm, Mathematica Policy Research, Inc., 600 Maryland Avenue, S.W., Suite 550, Washington, D.C. USA. 20024-2512, Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA USA 19104-6298.

where J is the number of racial (more generally, demographic) groups, a dot indicates summation over an index, and T and C superscripts designate true and census population counts, respectively. For state A in F and N's "counterexample," this expression implies

$$\left| \frac{90}{1,100} \left(\frac{1,100}{1,100} - \frac{89}{90} \right) + \frac{1}{1,100} \left(\frac{1,100}{1,100} - \frac{2}{1} \right) \right| = 0$$

$$\succ \frac{21}{13,750} = \left| \frac{90}{1,100} \left(\frac{979}{1,000} - \frac{89}{90} \right) + \frac{1}{1,100} \left(\frac{121}{100} - \frac{2}{1} \right) \right|.$$

Therefore, the condition for improved accuracy from adjustment is violated for state A. Similarly, for state B, we get

$$\left| \frac{910}{1,100} \left(\frac{1,100}{1,100} - \frac{890}{910} \right) + \frac{99}{1,100} \left(\frac{1,100}{1,100} - \frac{119}{99} \right) \right| = 0$$

$$\succ \frac{21}{13,750} = \left| \frac{910}{1,100} \left(\frac{979}{1,000} - \frac{890}{910} \right) + \frac{99}{1,100} \left(\frac{121}{100} - \frac{119}{99} \right) \right|.$$

Again, the condition for improved accuracy from adjustment is violated.

F and N's "counterexample" says nothing about our second analytical result. However, it is useful for numerically illustrating our third and clearly most important result. According to that result, when blacks are most heavily undercounted where they are *least* prevalent and whites are most heavily undercounted where they are *most* prevalent, synthetic adjustment *may* not improve the accuracy of the proportionate distribution. In F and N's example, state A has a higher black undercount than state B (50% versus 17%) but proportionately fewer blacks (2% versus 12%). State A has a higher white undercount (smaller overcount) than state B (-1% versus -2%) and proportionately more whites (98% versus 88%). Therefore, F and N's finding that the adjusted estimates in their example are less accurate overall than the census estimates, although not guaranteed, is not surprising in light of our third analytical result.

F and N's critique of our algebraic analysis of the effects of adjustment is based on a highly selective reading of our article that misrepresents our findings. F and N's criticism of our second analytical result is wrong as is their characterization of that result as central to our article. Our third analytical result is by far more important. It helps to expose those conditions on which adjustment's success or failure depends. Based on available empirical evidence cited below, the conditions of F and N's numerical example did not prevail in 1980, and our result suggests that synthetic adjustment would have improved the accuracy of the geographic distribution.

3. Simulation Results

As noted before, the purpose of our simulations was to answer several questions pertinent to synthetic adjustment and its effects on the accuracy of population estimates. The central questions addressed in our article were:

- How often would synthetic adjustment improve the accuracy of population estimates?
- How much would synthetic adjustment typically improve the accuracy of population estimates?

- Do the effects of synthetic adjustment on accuracy depend on how much census coverage varies from state to state?
- Do the effects of synthetic adjustment on accuracy depend on how well we measure national undercounts?

F and N focus on the second question. For the most part, we agree that the average magnitude of improvement in accuracy from synthetic adjustment is modest if our conservative assumptions about the state of the nation pertain. Under Case 22 of Scenario I, which probably exaggerates interstate variations in census coverage but is presented in S and P as our “moderate” variation case, the average reduction in the weighted sum of squared errors is just 8% while the average reduction in the unweighted sum of absolute errors is only 4%. It is important to understand, however, that larger improvements could be realized, as suggested by our third analytical result presented in S and P. The gains in accuracy would be somewhat greater, for example, if Hispanics had the same national undercount as blacks and were included with blacks instead of whites. In that case, the average reduction in the weighted sum of squared errors would be over 12%. The gains in accuracy would also be greater if black undercounts were higher in states with proportionately more blacks. We will return to this point shortly. Of course, improvements from synthetic adjustment might be smaller if there were substantial errors in measuring undercounts, although as we showed in S and P, the effects of measurement error are generally small.

What is easily forgotten in assessing the average gain in accuracy is the likelihood of realizing some gain, large or small. F and N are guilty of this oversight. Under the assumptions of Case 22, Scenario I, the likelihood of a gain in accuracy, according to the weighted sum of squared errors criterion, is 84%. We are impressed by this finding. Some improvement, perhaps only modest, is highly likely.

This result and the result on the average magnitude of improvement raise critical questions. What is the implication of the average improvement being “only modest”? Does the average improvement have to be overwhelming to justify adjustment? Put differently, should adjusted estimates be held to a higher standard than census estimates? The secretary of commerce imposed a higher standard in making the 1990 adjustment decision. How would the Census Bureau’s coverage improvement and imputation procedures fare by an equally high standard? We suspect that some would not fare well, having almost certainly exacerbated rather than ameliorated the differential undercount. Finally, would adjustment be recommended if it did little to improve accuracy but reduced systematic inequity? We will return to this last question in Section 4.

F and N answer these questions – which, by and large, do not have statistical answers – only implicitly, if at all. They suggest, however, that adjustment might be attractive (its estimates “will be good”) if the assumptions of our paper hold, the issue to which we now turn.

F and N wrongly characterize both the synthetic method and our simulation model. The underlying assumption of the synthetic method is not that there is no systematic geographic variation in undercounts for a given race but that there is no variation at all. Our simulation model shows how synthetic adjustment performs when this synthetic assumption is violated. We considered cases of extreme, albeit nonsystematic, interstate variation in undercounts by race, as well as cases with more moderate random variation. We did not construct true populations “on the basis of the synthetic assumption,” and our “definition of truth” did not “favor synthetic adjustment.” As we showed analytically in S and P, synthetic adjustment would have been favored by assuming a positive association between the black undercount and the prevalence of blacks or a negative association between the white undercount and the prevalence of whites. (A precise statement of the result is contained in the appendix to S and P.)

For purely illustrative purposes, we assumed in a new round of simulations that white undercounts are generated according to the assumptions of Scenario I, Case 22 in S and P but that the expected black undercount rises with the state's proportion black such that the black undercount is 2.0% when the proportion black is 11.7% (the national proportion black in 1980 according to the census) and 5.2% when the proportion black is 20.0%. Under those conditions, which we do not claim to be realistic although they preserve the average simulated differential in national undercounts, synthetic adjustment improves the accuracy of the proportionate distribution according to the weighted sum of squared errors criterion in all 1,000 iterations. The average reduction in the weighted sum of squared errors is over 17%, despite extreme variation in state total undercounts.

Do the assumptions of S and P pertaining to interstate variations in undercounts hold? Probably not. Although one of our purposes was to simulate a wide range of circumstances, it is very likely that our assumptions tended to put adjustment at a disadvantage.

Did we "offer no evidence" on the matter of geographic variation, as F and N claim? No, although admittedly there was not a wealth of information available. For judging our assumptions and their implications, there are two relevant empirical issues: whether variation is systematic or random and the extent of variation. We addressed both in S and P.

As we noted in S and P, according to the 1980 PEP blacks are hardest to count where they comprise large proportions of the population. In contrast, there is essentially no relationship between the white undercount and the relative prevalence of whites. (Ericksen and Kadane 1983). These conclusions are based on broad categories measuring racial composition and data for Standard Metropolitan Statistical Areas and state remainders, not state-level data. The only published undercount estimates by state and race are the "Developmental Estimates" for 1970. Although seriously flawed, based on heroic assumptions about internal migration (Wolter 1987), those estimates imply a direct relationship between the black undercount and the prevalence of blacks and a weak inverse relationship between the white undercount and the prevalence of whites. By ignoring either pattern of covariation, the simulations in S and P tend to *understate* the gains in accuracy from synthetic adjustment.

Since writing S and P, we have obtained unpublished state population and undercount estimates by race from the 1980 PEP. Because the raw black undercount estimates are imprecise for several states, it is not clear whether blacks are hardest to count – at the state level – where they are most prevalent. For whites, although there is evidence of a direct, rather than an inverse, association between their prevalence and the undercount, we believe that this is attributable to the inclusion of Hispanics in the white population and, to a much smaller degree, to the relatively heavy reliance on the conventional method of enumeration in a few predominately white states in the western U.S. Indeed, we find that if the true 1980 population followed the pattern of either the Series 2-9 or 10-8 estimates, a synthetic adjustment for the differential between the undercount of blacks and Hispanics and the undercount of all other persons would almost certainly have improved accuracy.

The available empirical evidence generally suggests that geographic variations are, if not random, systematic with a pattern that would enhance the gains in accuracy from synthetic adjustment. It seems unlikely that there is a strong inverse association across states between the black undercount and the prevalence of blacks or a strong direct association between the white undercount and the prevalence of whites. (Even if one or both of these patterns existed, adjusted estimates might still be more accurate, as we showed in S and P.) Thus, our assumption of randomness in S and P was probably conservative, working against synthetic adjustment.

Our assumptions about the extent of interstate variations in undercounts were also probably conservative, as we discussed in S and P with reference to the 1970 Developmental Estimates. Due to substantial sampling errors for many states, the unpublished state undercount estimates by race from the 1980 PEP do not reliably reveal how much black undercounts vary across states. The variance in black undercounts calculated across all 51 states is 0.0128 for Series 2-9, twice the highest assumed value in our simulations. The variance falls to 0.0036, not even midway between the moderate and high variances simulated, when New Hampshire (black undercount equal to -60%) and Vermont (black undercount equal to -24%) are excluded. (For Series 10-8, the interstate variance is nearly equal to the moderate value simulated in S and P, if three states with extreme (and highly unreliable) undercounts – less than -20% or greater than 20% – are excluded.) Raw estimates of state undercounts for whites from the 1980 PEP are far more precise. The interstate variances for Series 2-9 and 10-8 are just slightly below the moderate value simulated. The gains in accuracy under S and P's Scenario I, Case 12 (high variation among black undercounts and moderate variation among white undercounts) differ little in frequency or magnitude from the gains under Case 22 (moderate variation for both black and white undercounts), where improvements are highly likely.

From published PEP estimates for 1980, we can only calculate variances in total state undercounts, not differentiated by race. The largest interstate variance among the 12 published PEP Series is 0.00034, slightly less than the average simulated variance for our moderate variation case (Case 22). For Case 32 (low variation among black undercounts and moderate variation among white undercounts), the average simulated variance is 0.00031, about equal to the interstate variance for PEP Series 2-9, which is favored by Ericksen, Kadane and Tukey (1989) and is the median variance across the 8 PEP Series remaining after excluding 10-8, 14-8, 14-9, and 14-20. Synthetic adjustment reduces the weighted sum of squared errors by about 12% on average under Case 32, compared to 8% for Case 22. Case 23 (moderate variation among black undercounts and low variation among white undercounts) implies an average simulated variance only slightly greater than the variance for PEP Series 10-8, F and N's "favorite." Under the conditions of Case 23, synthetic adjustment reduces the weighted sum of squared errors by 19% on average. Adjusted estimates are more accurate over 92% of the time according to that error criterion. Are such improvements "only modest"?

4. Accuracy and Equity

We have argued before, in S and P and in Schirm (1991), that the foremost concern of statisticians and demographers should be the accuracy of population estimates. Yet, in a single-minded pursuit of statistical accuracy, it is easy to forget considerations of political equity.

A more accurate population distribution is probably more equitable, in general. However, this does not imply that two equally accurate distributions are equally equitable. Although adjustment may do little to improve overall accuracy in a particular year, it may reduce or remove certain systematic errors and systematic inequity, errors and inequity associated with race.

An example, obtained from our simulations, is displayed in Table 1. The implied black and white national undercounts are 5.2% and -1.1% . The adjusted population estimates in Table 1 were obtained using these figures and the synthetic method.

As will become clear, it is hard to draw a sharp distinction between accuracy and equity. For this discussion, we assume that accuracy is narrowly defined in terms of the proportionate geographic distribution.

Table 1
A Numerical Example: Population Counts (1,000s)

State	True		Census		Adjusted	
	Black	White	Black	White	Black	White
Alabama	1,060	2,849	996	2,898	1,051	2,866
Alaska	14	375	14	388	15	384
Arizona	83	2,606	75	2,643	79	2,614
Arkansas	414	1,916	374	1,912	395	1,891
California	1,758	22,105	1,819	21,849	1,919	21,607
Colorado	106	2,719	102	2,788	108	2,757
Connecticut	226	2,875	217	2,891	229	2,859
Delaware	101	507	96	498	101	492
District of Columbia	442	181	449	189	474	187
Florida	1,415	8,290	1,343	8,403	1,417	8,310
Georgia	1,642	3,910	1,465	3,998	1,546	3,954
Hawaii	18	925	17	948	18	938
Idaho	3	936	3	941	3	931
Illinois	2,014	9,677	1,675	9,752	1,768	9,644
Indiana	443	4,775	415	5,075	438	5,019
Iowa	43	2,809	42	2,872	44	2,840
Kansas	135	2,284	126	2,238	133	2,213
Kentucky	254	3,327	259	3,402	273	3,364
Louisiana	1,239	2,856	1,238	2,968	1,306	2,935
Maine	3	1,135	3	1,122	3	1,110
Maryland	1,026	3,159	958	3,259	1,011	3,223
Massachusetts	230	5,223	221	5,516	233	5,455
Michigan	1,272	7,990	1,199	8,063	1,265	7,974
Minnesota	56	4,037	53	4,023	56	3,978
Mississippi	891	1,629	887	1,634	936	1,616
Missouri	535	4,329	514	4,403	542	4,354
Montana	2	761	2	785	2	776
Nebraska	51	1,521	48	1,522	51	1,505
Nevada	52	760	51	749	54	741
New Hampshire	4	911	4	917	4	907
New Jersey	1,071	6,237	925	6,440	976	6,369
New Mexico	27	1,264	24	1,279	25	1,265
New York	2,397	14,891	2,402	15,156	2,535	14,988
North Carolina	1,387	4,443	1,319	4,563	1,392	4,512
North Dakota	3	622	3	650	3	643
Ohio	1,112	9,769	1,077	9,721	1,136	9,613
Oklahoma	208	2,819	205	2,820	216	2,789
Oregon	39	2,602	37	2,596	39	2,567
Pennsylvania	1,177	10,750	1,047	10,817	1,105	10,697
Rhode Island	29	923	28	919	30	909
South Carolina	962	2,182	949	2,173	1,001	2,149
South Dakota	2	678	2	689	2	681
Tennessee	754	3,764	726	3,865	766	3,822
Texas	1,752	12,421	1,710	12,519	1,804	12,380
Utah	10	1,435	9	1,452	9	1,436
Vermont	1	523	1	510	1	504
Virginia	1,116	4,356	1,009	4,338	1,065	4,290
Washington	109	3,867	106	4,026	112	3,981
West Virginia	69	1,877	65	1,885	69	1,864
Wisconsin	198	4,588	183	4,523	193	4,473
Wyoming	3	448	3	467	3	462
Total	27,958	197,836	26,495	200,054	27,956	197,838

Note: "White" includes all nonblacks.

Table 2
A Numerical Example: Congressional Apportionments

State	Number of House Seats		
	True	Census	Adjusted
Alabama	8	7	8
Alaska	1	1	1
Arizona	5	5	5
Arkansas	4	4	4
California	46	45	45
Colorado	5	6	5
Connecticut	6	6	6
Delaware	1	1	1
District of Columbia	0	0	0
Florida	19	19	19
Georgia	11	10	11
Hawaii	2	2	2
Idaho	2	2	2
Illinois	22	22	22
Indiana	10	10	10
Iowa	5	6	6
Kansas	5	5	5
Kentucky	7	7	7
Louisiana	8	8	8
Maine	2	2	2
Maryland	8	8	8
Massachusetts	10	11	11
Michigan	18	18	18
Minnesota	8	8	8
Mississippi	5	5	5
Missouri	9	9	9
Montana	2	2	2
Nebraska	3	3	3
Nevada	2	2	2
New Hampshire	2	2	2
New Jersey	14	14	14
New Mexico	3	3	3
New York	33	34	33
North Carolina	11	11	11
North Dakota	1	1	1
Ohio	21	21	21
Oklahoma	6	6	6
Oregon	5	5	5
Pennsylvania	23	23	23
Rhode Island	2	2	2
South Carolina	6	6	6
South Dakota	1	1	1
Tennessee	9	9	9
Texas	27	27	27
Utah	3	3	3
Vermont	1	1	1
Virginia	11	10	10
Washington	8	8	8
West Virginia	4	4	4
Wisconsin	9	9	9
Wyoming	1	1	1
Total	435	435	435

Although the census and adjusted distributions in Table 1 are equally accurate, according to a weighted sum of squared errors criterion, the adjusted estimates more accurately reflect the racial distribution at the national level and are more equitable. (The census geographic distribution is slightly more accurate according to a sum of absolute errors standard.) The true and adjusted figures imply that 12.4% of the U.S. population is black. According to the census, only 11.7% of the population is black, a serious inequity.

The equity gains from adjustment are made more concrete by the implied congressional apportionments shown in Table 2. Both the census and adjusted estimates allocate one too many seats to Iowa and Massachusetts and one too few to California and Virginia. However, the census estimates also allocate one too many seats to Colorado and New York and one too few to Alabama and Georgia, whereas the adjusted estimates allocate the correct numbers of seats to these states. Based on the census figures, Alabama and Georgia are denied representation because of their high proportions black and the differential undercount to which blacks are subject. Adjustment substantially improves equality of representation. Based on the true population figures and the census apportionment, there are 471,000 persons per representative in Colorado, 508,000 in New York, 555,000 in Georgia, and 558,000 in Alabama. Adjustment narrows the differences, with 565,000 persons per representative in Colorado, 524,000 in New York, 505,000 in Georgia, and 489,000 in Alabama. For the four states combined, there are 519,000 persons for each of the 57 representatives. (For the entire U.S., there are 519,000 persons for each of the 435 representatives.) Adjustment reduces the (unweighted) root mean square deviation from this average by over 21%. (The reduction is between 20% and 21% when deviations are weighted by the number of persons per representative according to the true population figures.) The equity gain from adjustment is also clearly revealed by the weighted average of persons per representative calculated across all 50 states. When weighted by the proportion of the national black population (exclusive of the District of Columbia) living in the state, the true average number of persons per representative is 518,000. If Congress is apportioned according to the census estimates, the average is 524,000. Synthetic adjustment removes most of this racial inequity. The average number of persons per representative is 520,000 when House seats are allocated according to the adjusted estimates. Although there are surely still other ways to measure inequality of representation, it is hard to imagine a reasonable alternative that would not show adjustment reducing the racial inequity attributable to differential census undercounting. The gain in equity in this example is achieved despite no gain in accuracy of the proportionate distribution across states.

The errors in the census are systematic. After adjustment, the remaining errors may not be truly random, and so long as there are errors, there will be inequity. However, the source of those errors would be far less offensive than race.

5. Discussion

In criticizing S and P and Ericksen, Kadane and Tukey (1989), F and N emphasize the role of assumptions underlying adjusted estimates. Their view, however, is extreme, counterproductive, and fundamentally flawed.

Although it is reasonable – and necessary – to ask whether assumptions matter, assumptions do not have to be exactly true as F and N imply. Moreover, proponents of adjustment should have to defend it against only reasonable alternative assumptions. F and N seem to believe that almost any alternative is fair game. As in assessing the magnitude of improvement, they require adjustment to bear a heavy burden of proof with no scientific justification. Nonetheless, as our simulations show, synthetic adjustment improves accuracy even under extremely unfavorable – and probably unreasonable – assumptions.

This raises another important point of disagreement between us and F and N. Not all assumptions have to imply precisely the same estimates if all adjusted estimates based on reasonable assumptions are more accurate than census estimates. Unless equally plausible assumptions have very different implications, we should not reject the better for failure to find the best. We should not settle for census estimates that are less accurate.

F and N offer nothing to suggest that census estimates are more accurate than adjusted estimates. The legal findings on which they rely are no basis for a scientific argument. Moreover, despite Schirm's (1991) finding that the judgmental decisions made in producing census estimates can affect congressional apportionment, F and N fail to scrutinize census procedures and the underlying assumptions. Do they find plausible the census "assumption" that when the final estimates are released, everyone everywhere has been counted in exactly the correct location? If not, do they have any constructive suggestions for improving the accuracy of population estimates? Unfortunately, their critical commentary on those suggestions that have been offered is seriously flawed by misrepresentation and distortion and offers nothing constructive.

"Should we have adjusted the census of 1980?" as F and N ask. Maybe, maybe not. Although it is subject to debate, we may not have known enough about the likely effects of adjustment or been technically and operationally prepared to undertake an adjustment at the time a decision had to be made. Would adjustment have improved accuracy in 1980? We cannot answer with certainty because the true population is inherently unknowable and anomalies cannot be entirely ruled out. With that qualification, the answer is "very likely".

Acknowledgements

The authors thank Gene Ericksen for providing unpublished estimates from the 1980 PEP. The views expressed are those of the authors alone.

ADDITIONAL REFERENCES

- ERICKSEN, E.P., and KADANE, J.B. (1983). Using the 1980 Census as a population standard. *Proceedings of the Social Statistics Section, American Statistical Association*, 474-479.
- WOLTER, K.M. (1987). Comment on Schirm and Preston (1987). *Journal of the American Statistical Association*, 82, 978-980.

COMMENTS

J.A. HARTIGAN¹**The Adjustment Controversy**

Each ten years the United States Census prepares a list, or *enumeration* of names and addresses of persons resident in the United States. The list is subject to error in that persons may be omitted from the list, or erroneously included on it. In the 1990 census, Erickson *et al.* (1991) estimated there to be 13 million erroneous enumerations and 17 million omissions. Even if these estimates are off by a factor of two, it appears that the enumeration needs some adjustment.

Freedman and Navidi discuss statistical evidence presented in a law suit intended to force the Bureau of the Census to adjust the 1980 enumeration. The origin of the law suit is the *differential undercount* between races. The undercount is perhaps 5% for Blacks and Hispanics, and 1% for Others. Since the undercount is greater for minorities, those localities with larger fractions of minorities press for an adjustment in the census figures that would adjust for the undercount. The undercount has been established by Demographic Analysis (counting births, deaths, emigration, and immigration by race, sex, and age) in censuses since 1940, and by Post Enumeration Surveys, surveys to obtain a more accurate count in a sample of the population, since 1970. The size of the undercount is an important point of dispute, since it makes more sense to estimate and correct for a large differential undercount than for a small one. Freedman and Navidi concede that there may be a differential undercount, but assert, at least for the 1980 census, that the undercount is not sufficiently well estimated in different localities to make adjustments feasible. Freedman and Navidi are concerned mainly to criticize proposed techniques for doing the adjustment; what are their own estimates of the undercount? For example, do they agree that the national undercount is as high as 5% for Blacks and Hispanics compared to 1% for others? I will argue later that if the differential undercount is that high, a synthetic adjustment (each minority person weighted 1.05, each majority person weighted 1.01) will probably improve estimates of state population shares.

In 1980, the Bureau conducted a Post Enumeration Program which it intended to use in adjustment. The bureau decided not to adjust, on the grounds that the PEP estimates were not sufficiently accurate or reliable to give improved counts in small localities. This paper reprises Freedman's testimony in the court case which followed, in which the court's decision supported the Bureau of Census. It may be of interest to report some of the later developments, which show that the issues raised in the present paper are still very much alive. In the 1980's the Bureau planned a more substantial Post Enumeration Survey for the 1990 Census. A dress-rehearsal PES was run in 1988. Some 20 evaluation studies to handle various types of error in the PES were planned and carried out after the 1990 Census. In 1988, the Secretary of Commerce announced that there would be no adjustment of the 1990 census. The government was sued by various localities with high fractions of minorities. The secretary then agreed, on 17 July 1989, to continue planning for the PES, to appoint a committee of 8 experts who would advise the secretary on the feasibility of adjustment, and to publish a set of guidelines under which the Census enumeration would be adjusted or not. The external committee met frequently with Census officials, and advised them on planning, execution and analysis of the PES. The evaluative analyses were carried out by the bureau, and on 21 June, 1991, the steering committee in the census, with some dissent, recommended to the secretary that the census be adjusted.

¹ J.A. Hartigan, Department of Statistics, Yale University, New Haven, CT USA 06520-2179.

The recommendation of the committee was considerably weakened a few days later when an earlier analysis was found to be in error. The external committee divided into two groups of four. The first of these, Ericksen, Estrada, Tukey, and Wolter, with the aid of many consultants, wrote an extensive report that found many defects with the original enumeration, and strongly urged adjustment. The second group of four, Kruskal, McGehee, Tarrance, and Wachter, are as strongly opposed to adjustment. Wachter, with the help of some consultants, offers alternative statistical analysis of the PES that suggest the range of plausible adjustments is so wide as to have quite different effects on reapportionment and other distributive requirements of the Census figures. The secretary decided that the statistical foundation for adjustment was inadequate and recommended against adjustment. The Department of Commerce was sued by the same localities that sued in 1980. The 1980 court case is thus being replayed after the 1990 census.

Synthetic Adjustment

A simple *synthetic* scheme is to multiply each minority person actually enumerated by 1.05 and each majority person actually enumerated by 1.01. I agree with Freedman and Navidi's rejection of Schirm and Preston's (1987) analytic argument.

What about the following analytic argument? Suppose that national undercounts are correctly estimated, but the undercounts differ over states; when does the synthetic adjustment improve the estimate of a state's proportion of the national population? The answer is, if the synthetic adjustment is an overadjustment for a particular state, it is closer to the true proportion than the enumerated proportion if and only if the minority fraction in the state is less than the national minority fraction; conversely, if the adjustment is an underadjustment for a particular state, it is closer to the true proportion than the enumerated proportion if and only if the minority fraction in the state is greater than the national minority fraction. It is plausible to expect the undercount for minorities and non-minorities in high minority states to be higher than in low minority states, which would cause the synthetic adjustment to be an under-adjustment, but nevertheless to be an improvement on the enumerated proportion.

National undercount rates of 5% and 1% are supported by historical evidence from the Bureau, both by demographic analysis and post enumeration surveys, Tables 1 and 2. In a 5-1 adjustment, we multiply non-minority enumerations by 1.01 and minority enumerations by 1.05. Will this improve apportionment of Congressional seats to the various States?

The actual minority and non-minority populations in the different States are unknown. We are comparing the two estimates of the State populations based on the unadjusted and adjusted Census. The census will do best when the States with high minority populations actually have a low differential undercount; the 5-1 adjustment will then overshoot the true proportions for those states. Correspondingly, if the States with low minority populations actually have a high differential undercount, then the 5-1 adjustment will undershoot the true proportions. This tells us how to construct a best case for the census, and a worst case for the adjustment.

I will ignore variations in the non-minority undercount between States, as these should have a minor affect on the overall proportions; I will suppose that all States have a non-minority undercount of 1%. Suppose that the true overall minority undercount is 5%. Suppose this undercount might vary from 3% to 7% in the different States. I assign 3% undercounts to high-minority states, and 7% undercounts to low-minority states, with the division between high and low minority being decided so that the overall minority undercount is 5%. This assumption of true undercounts makes the census look best. The calculation is done for a range of choices of overall undercount and variations across states.

Table 1
Historical Estimates of the Amount and Percent of Net Undercount by Race,
as Measured by Demographic Analysis
(Report dated 21 June 1991 from the Bureau of Census undercount steering committee)

	1940	1950	1960	1970	1980	1990
Total	5.4	4.1	3.1	2.7	1.2	1.8
Black	8.4	7.5	6.6	6.5	4.5	5.7
Non-black	5.0	3.8	2.7	2.2	0.8	1.3

Table 2
Undercount Estimated by the Post Enumeration Survey and Demographic
Analysis in the 1990 Census, by Age, Race and Sex

	Black				Non-black			
	Male		Female		Male		Female	
	PES	DA	PES	DA	PES	DA	PES	DA
	5.4	8.5	4.3	3.0	2.0	2.0	1.4	0.6
0-9	8.0	8.2	7.8	7.8	3.3	2.7	3.4	2.8
10-19	4.0	2.0	4.0	2.2	1.2	-1.0	1.8	-0.5
20-29	6.4	9.4	6.8	3.8	5.0	2.1	3.8	0.9
30-44	5.9	12.4	3.9	2.5	2.2	2.7	1.4	0.1
45-64	3.2	11.7	1.3	0.5	0.4	2.8	-0.5	0.4
65+	1.0	3.0	-0.3	-1.3	-0.9	1.4	-1.1	0.4

The census and adjusted estimates are compared by computing the number of congressional seats that are wrongly apportioned by the two estimates of population. The number of seats allocated to a State with 7.2% of the population is 435×7.2 suitably rounded. The rounding makes the actual apportionment a rather poor measuring rod for comparing two methods, because the misapportionment is usually only 1 or 2 seats. Instead, I will use fractional misapportionment, which is half the sum of absolute differences between the estimated and true proportions, multiplied by 435.

It can be seen from Table 3, that the break – even point for census versus adjusted occurs when the true overall minority rate is 3%; we expect this, because then the true differential undercount is 2%, half-way between the 0 rate implied by the census, and the 4% rate assumed by the adjustment. For higher overall minority undercounts, the census does better only when there is a big range of variation across states, and the states with high minority populations happen to have low undercount rates. For example, if the overall rate is 4%, the census achieves 0.8 misapportionment against 1.2 for the adjustment, provided that all the high-minority states have a 2% undercount, and all the low-minority states have a 6% undercount. If the overall rate is 5%, the census achieves 0.8 versus 1.0, only if the high-minority states have a 3% undercount and the low-minority states have a 7% undercount. If the overall rate is greater than 5%, the census is better than the adjusted for no combination of undercount rates in the states having a range of 4% or less.

Table 3

Comparison of Fractional Misapportionment for the Census and a 5-1 Adjustment for a Range of Overall Minority Undercount Rates, with Varying Undercount Rates for the States (The majority undercount rate is fixed at 1%; the minority populations in each state are estimated from the U.S. statistical abstract, 1989)

Overall minority undercount	Minority undercount for low-minority states	Minority undercount for high-minority states	Census fractional misapportionment	5-1 fractional misapportionment
2	2	2	0.2	0.7
2	3	1	0.4	1.1
2	1	3	0.7	0.6
3	3	3	0.5	0.5
3	4	2	0.4	0.9
3	2	4	0.9	0.4
4	4	4	0.7	0.2
4	5	3	0.6	0.7
4	3	5	1.1	0.4
4	6	2	0.8	1.2
4	2	6	1.6	0.9
5	5	5	0.9	0.0
5	6	4	0.8	0.5
5	4	6	1.3	0.5
5	7	3	0.8	1.0
5	3	7	1.8	1.0
6	6	6	1.2	0.2
6	7	5	1.0	0.4
6	5	7	1.6	0.7
6	8	4	1.0	0.9
6	4	8	2.0	1.2
7	7	7	1.4	0.5
7	8	6	1.2	0.4
7	6	8	1.8	0.9
7	9	5	1.2	0.8
7	5	9	2.2	1.4

Table 3 suggests that the overall rates would need to be 3% or less to make this crude 5-1 rule less accurate than the census for apportionment. The 1990 PEP-based 95% 'confidence intervals' for the overall minority rate are 4.3 to 5.7; this range seems overoptimistically narrow, but even if we doubled the quoted margins of error, the interval is 3.6 to 6.4; if the true value lies in this range, then the 5-1 rule will still beat the census.

PEP and PES Based Adjustments

The 1980 PEP survey and the 1990 PES survey are designed to refine a synthetic adjustment by estimating different undercount rates in different localities. Freedman and Navidi are skeptical about the regression used to smooth the estimates, questioning independence,

homogeneity of variance, and the reliability of the selection procedures for including variables in the model. I was not persuaded by their examples of how different sets of variables could easily have been selected for inclusion in the model; after all, if the predicting variables are highly correlated, quite different subsets can produce pretty much the same prediction. Thus the fact that different variables were selected does not indicate the smoothed estimates would be very different. Indeed, their table 10 indicates that two variables, percent minority, and percent conventionally enumerated appeared in nearly all equations. I would suspect that the *assumptions* of the regression can not be easily defended, but that the *results* of the regression are reasonable, except perhaps in producing lower standard errors than are justified by the probable lack of independence.

Reduction in sampling variance by regression-based smoothing procedures is not likely to make much difference to estimates in large localities such as States. There, the aggregation of different PEP or PES estimates is already doing as much smoothing as is needed, and the questionable regression assumptions can be avoided. On the other hand, the regression smoothing probably is needed if results are to be projected to small localities.

I agree with Freedman and Navidi that missing data procedures and bias assessment in the post surveys are the key to evaluating the adjusted estimates. Correct handling of missing data, and assessing bias, requires an intimate understanding of survey procedures. Personal judgements by the professionals most closely involved will dominate the conclusions. A healthy skepticism about any resulting 'standard errors' or 'confidence intervals' is justified.

I suggest that the the right loss function to evaluate accuracy in apportionment is not squared error, which is statistically convenient for combining variances and squared biases, nor estimates of the numbers of states or localities that are better estimated by the adjustment. For apportionment, the loss function should be the sum of absolute differences between estimated and true proportions in the different states, because this represents the numbers of people actually misallocated by the estimates, and corresponds at the state level to the number of misapportioned seats.

Although state proportions are of primary interest, let's look at the state populations first. If the true undercount rate in a state is 2%, then the census is better than the estimate just when the estimated undercount rate is less than 0% or greater than 4%. This occurs with probability 50% when the standard error of the estimate is about 3%; thus even a quite inaccurate estimate of undercount is enough to give the adjusted estimate the edge. The census has the same expected difference from truth as the estimate when the standard error is 2.2, and the same expected square difference when the standard error is 2. Out of all this comes the simple rule, that if the true rate is 2%, you do better than the census if you can estimate the true rate with standard error 2%. When estimating population proportions, rather than populations, the relevant computations are on the differential undercounts for the various states, the difference between the undercount for the state and the nationwide undercount, (not the difference between the races); thus adjustments do better than the census in those states where the true difference between the state undercount rate and the nationwide undercount rate exceeds the standard error of the estimated difference.

Under this rule, and accepting the bureau's 1990 estimates of undercount rates and margins of error based on the Post Enumeration Survey, the enumeration is estimated to do better in 24 out of 50 states in estimating proportions. Note however, that the overall estimated loss is quite a bit better for the adjustment than the census, because the states with large (plus or minus) differential undercount rates are estimated better by the adjustment; when the census does better, it does just a little better; when the adjustment does better it often does quite a lot better. Thus the fact that 24 out of 50 states are not estimated to be improved by adjustment

should not cause too much excitement; it just means that a lot of states have an estimated undercount rate that is pretty close to the national average, and there is no advantage, and not much difference, in adjusting them. We continue to estimate that substantially more people are correctly allocated by the adjusted figures than the enumeration. Table 4 gives some error estimates when the census PES based figures are incorrect in various ways.

The bureau has produced a number of estimated undercounts, with margins of error, in the various states. I use the ‘selected PES method’ (called PES from now on) in the report of the Undercount Steering Committee, 21 June 1991. Now there are two popes, the enumeration, and the PES figures. Which is correct? Well, you need a third pope, an infallible one, to decide.

We don’t have the third pope. The various follow up evaluations of the PES, the total error model, the loss function analyses, the robustness analyses, are all attempts to feel out what the third pope might decide, but an attitude of skepticism and caution is necessary in believing the decisions of the fictional third pope. In particular, the bureau’s ‘true population’ estimates are all variations on the PES estimates, accepting the basic accuracy and feasibility of the PES, and so most unlikely to find the PES at fault compared to the enumeration. The PES can only be found inferior by some method that is not so closely linked to it. Demographic analysis is by no means in complete agreement with the PES, and provides only national information, but on the whole, it supports the PES rather than the enumeration.

I have done some sensitivity analyses to evaluate how far the PES estimates and margins of error would have to be in error for the census to look competitive. The calculations use the PES state undercount estimates with various multipliers, the PES margins of error with various multipliers, and assume that the true state figures are sampled from normal distributions with the multiplied PES state undercounts and margins of error. I take the different state truths to be independent, which is surely far from correct. The independence will not seriously affect the individual state proportions though, so the average misapportionment of the census and the PES won’t be much affected; the variability of the difference will be underestimated.

The results in Table 4 show that the PES estimates have to be in substantial error before the CENSUS starts to be competitive. Accepting the PES rates and margins of error, the CENSUS misapportions 4 seats, the PES 1. If the PES overcount rates are halved, with the margins of error remaining fixed, then the misapportionment rate for the census is 2.5 seats, and for the PES 1.5 seats, and the census will be better in about 40% of the true cases.

However this analysis is in line with the loss function analyses in that it takes the PES as its starting point.

Table 4
Misapportionment of the Enumeration and PES-adjustment, when the True Figures
are in Accordance with the PES-undercount Rates and Margins of Error,
with Various Multipliers for the Undercounts and Error Margins
(Based on 100 simulated true counts.)

Multiplier for pes undercount in each state	Multiplier for margin of error in each state	Census misapportioned seats	Adjusted misapportioned seats	Standard deviation of difference
1	1	3.8	1.1	1.2
1	2	3.7	2.0	1.3
0.5	0.5	2.8	1.2	1.3
0.5	1.0	2.5	1.6	1.4
0.75	0.75	3.3	0.9	1.3
0.75	1.50	3.3	1.6	1.3

I wonder if it might not be useful to make a distinction between the *enumeration*, that lists names and addresses, the *count*, that counts the number of people in various localities according to the list, and *estimates*, using statistical procedures based on various sources of information such as demographics and supplementary lists. This would be perhaps politically divisive, since interested parties would wish to allocate according to the figures most favourable to themselves. There would be the danger that the census professionals, with several estimates available, would be subject to political pressure to choose estimates favourable to one or other group. If we want to estimate populations accurately though, it is a good principle to base the estimates on several mutually supporting surveys rather than a single one. The danger of relying on a single list outweighs the dangers and difficulties in combining information from different sources. For one thing, the only way to find out how accurate a survey is to compare it to another survey in one form or another. It should be noted that even in the 'unadjusted' census, population estimates are not simple counts off an enumerated list. Individuals known to be fictitious are included in the count by various kinds of imputation procedures that handle missing data.

Perhaps Freedman and Navidi are right, in asserting that the 1980 PEP figures were too unreliable to permit their use in adjusting the census figures. Perhaps they were right in saying that a nationwide synthetic adjustment is too crude, and has not been demonstrated to improve accuracy. Yet omissions and erroneous enumerations in the tens of millions suggest that some kind of adjustment might improve accuracy. There is plenty of room for improvement. We could misguess the existence or location of a few million people and still be competitive with the raw enumeration.

I have some questions for Freedman and Navidi.

- (1) Do they agree with these estimates of 13 million omissions and 17 million erroneous enumerations?
- (2) Is the nationwide differential undercount between Blacks and Whites 4%?
- (3) Is the PES a useful tool for assessing accuracy of the first census?
- (4) Should the PES follow-up sample be used to correct the first census, not only in the specific instances of erroneous enumerations and omissions discovered by comparing the surveys, but also by projecting differential undercounts discovered in the follow up sample to the whole census? If so, how?
- (5) If the PES is not good enough, how should the follow-up survey be designed so that it could be used to adjust the census?

COMMENT

T.P. SPEED¹

Freedman and Navidi ask "Should we have adjusted the census of 1980?" and answer no. I take this as meaning that they have yet to see compelling evidence that it could have been done well, not that they do not see a problem, and not that they think adjustment intrinsically undesirable. My interest in census adjustment was aroused about four years ago, shortly after I came to the U.S. I have read the papers by the main participants in the debate, and have recently had the opportunity to examine some block-level data from the 1990 Post Enumeration Survey. My conclusion is the same as that of Freedman and Navidi: there is simply no evidence to show that adjustment will work at the level proposed.

There are two features of the arguments for adjustments that I find particularly striking. Absolutely no use is ever made of "ground truth" data to demonstrate clearly that adjustments do improve upon the census. And no use is made of available data to justify the key assumptions on which adjustment methods are based.

In 1990 adjustment was to be at the census block level. This would have meant that on the basis of a sample of about 12,000 blocks, each of 6.5 million blocks would have been adjusted. Adjusting a census block count means adding or subtracting people with specific characteristics before further aggregation. This would have been done using procedures based on unverified and implausible assumptions concerning the undercount mechanism. The most important such assumption is that undercount rates are constant within 1,392 demographic subgroups of the population called poststrata, defined by region, race, sex, age and status as a home owner or renter. One such consists of all non-black male Hispanic renters aged 30-44 living in Los Angeles city, or in central cities in the Pacific Census Division (California, Oregon, Washington, Alaska, Hawaii). Another consists of all female owners aged 20-29 living in central cities of 250,000 or more, excluding New York City in the Mid-Atlantic Census Division (New York, New Jersey, Pennsylvania) who are not black, Hispanic, Asian or Pacific Islanders. The parallel with the regression models in the present paper is clear.

Examination of block-level data from the 1990 Post Enumeration Survey from sites in Detroit and Texas showed that the assumption of constancy of the undercount rates within 1,392 poststrata is no better supported than a quite different one: that the undercount is driven by blocks, and is constant across poststrata within blocks. This dual model would have led to different block-level adjustments. The analysis is difficult because the counts of people in the intersections of blocks and poststrata are quite small, heterogeneous, and mostly zero. Details can be found in Hengartner and Speed (1992).

Of course I do not know whether the poststrata-driven or the block-driven undercount model is better; we would need something like "ground-truth" data to answer that. But we can see that certain key assumptions concerning the 1990 undercount model are no better supported by available data than those of a quite different model. In my view, when changing assumptions changes the results, and when we have no way of telling which set of results is closer to the truth, then we have no business adjusting. This is the message I get from the present paper, and it is one I wholeheartedly support.

ADDITIONAL REFERENCE

HENGARTNER, N., and SPEED T.P. (1992). Assessing between-block heterogeneity within poststrata of the 1990 Post-Enumeration Survey. Submitted to *Journal of the American Statistical Association*.

¹ T.P. Speed, Department of Statistics, University of California, Berkeley, CA U.S.A. 94720.

COMMENT

EUGENE P. ERICKSEN and JOSEPH B. KADANE¹

“The Court (Judge Sprizzo): I take it your standard error should be a fixed statistical number which you then subtract from your results and you get what is left, basically which is supposed to measure the accuracy of what you are measuring?”

The Witness (David Freedman): I hate to argue with you, but it isn’t quite like that.”
(Cuomo v. Baldrige: 2629).

We welcome the opportunity to continue the debate with Freedman and Navidi (F and N). Although Judge Sprizzo’s decision is now more than 4 years old, the statistical issues are important ones and deserve continued attention. This is especially true because the final scientific judgements are best made by statisticians and demographers, rather than judges and politicians. In this article, Freedman and Navidi review their side of the adjustment controversy, explore some new arguments, and try to use Judge Sprizzo’s legal decision to support their scientific position. In this comment, we reexamine certain critical points, restating and clarifying our position where necessary in an effort to demonstrate how adjusting the 1980 Census would have made the data more accurate.

Our disagreements with Freedman and Navidi are fundamental, and we agree that they go to the heart of statistical inference. In their conclusion, F and N write “success of any of EKT’s proposed adjustments rides on unverified and implausible assumptions (p. 19).” To the contrary, we believe that our assumptions are realistic and verified by decades of census-taking knowledge, as we will argue below. For their part, F and N’s arguments boil down to little more than concern that some assumptions may not be true. To criticize a statistical argument however, it is necessary to do more than that. Assumptions are usually not true exactly – the relevant question is how far they are from being exactly true and what that means for the intended uses of the data. At a minimum, one must show that other assumptions, argued to be just as realistic, or more realistic, lead to substantially different conclusions. F and N do none of this. Moreover, although they concentrate upon the minor differences in various adjustment possibilities, they make no attempt to demonstrate that the adjustments would result in estimates with larger errors than the unadjusted census.

An important part of the disagreement concerns whether or not it is proper to use what we know about the census. F and N give no weight to evidence of greater census-taking problems in some areas than others, and give no credit to the fact that the PEP-measured omission rates and undercounts are higher in those areas with lower mailback rates, higher rates of missing data, and greater problems maintaining the specified long-form sampling rate on the census. Nor do they give any credence to the consistency of the racial differentials in undercount provided by demographic analysis for every census since 1940. This information is not relevant to them, and they are quick to criticize us whenever we rely upon “unverified” assumptions, no matter how realistic or warranted. They also do not explain what “verification” is to them.

At the same time, Freedman and Navidi were not able to make their own argument without reliance upon assertions which are either unverified or are based on the very PEP data which they criticize us for using. Here are some examples:

¹ Eugene P. Ericksen, Temple University, Philadelphia, PA. USA. and Joseph B. Kadane, Carnegie-Mellon University, Pittsburgh, PA USA 15213.

1. A small undercount is thought to remain in the census (p. 3).
2. The census also had a small amount of erroneous enumerations (p. 4).
3. The undercounts estimated by PEP are likely to be biased upward (p. 12).
4. The eruption of Mt. St. Helens caused correlated error between the original enumeration and the PEP (p. 12).
5. Missing data caused a bias in the PEP (p. 13).
6. Minority persons living in central cities are likely to behave differently from those in suburbs (p. 16).
7. The undercount in conventional areas was relatively high (p. 18).

F and N seem to believe that in the absence of substantial direct information about the quality of the PEP data that we should not adjust since different assumptions sometimes lead to different results. This argument, however, ignores the well documented errors in the census enumeration. We have provided extensive documentation, not just in the EKT article, but elsewhere (Erickson 1983; Erickson and Kadane 1985) of problems in the census, and others (Citro and Cohen 1985; U.S. Bureau of the Census 1985, 1986 and 1988) have found similar results. To us, the substantial evidence of census-taking problems, the geographic coincidence of census-taking problems with high PEP undercounts, and the consistency of the PEP series with each other and with the results of demographic analysis results provide ample assurance that the additional information derived from the PEP data could have been used to adjust – and improve the accuracy of – the 1980 Decennial Census. This summarizes our general point of view. In the sections that follow, we address some of Freedman and Navidi's specific arguments.

Do the Simple Adjustments Improve Upon the Census?

In their Section 2, F and N criticize our Table 5, in which we claim to show the general agreement of 14 different adjustment schemes. Each of them shifts population share from predominantly White areas outside of cities, where census-taking problems were low, to large central cities with substantial minority populations, where census-taking problems were great. F and N conclude: "The table does not show that any of the methods improve upon the census. It cannot, because there is no external standard against which to measure improvement" (p. 6). If what F and N mean is that the "true" population is unknowable, than their argument, of course, goes too far and no adjustment could ever meet their requirements.

In the EKT paper, we relied upon Schirm and Preston (1987) to show that a simple synthetic method (our Synthetic B) improved upon the census. Since they are also commenting upon Freedman and Navidi's article, we will not repeat their arguments. Given the improvement provided by Synthetic B, we would expect further improvement to be provided by more realistic assumptions, namely that minority populations would be more difficult to count in areas where census-taking problems are greater. These assumptions are consistent not only with PEP results, but with the result of a separate Census Bureau study of New York City which showed that omission rates were strongly and negatively correlated across district offices with mailback rates (Erickson and Kadane 1986).

Freedman and Navidi base their argument on the apparent differences in the adjusted distributions provided by the different PEP series. We do not believe this evidence to be pertinent, since we know that the eight "preferred PEP's," as well as the more reasonable Synthetic A, will be different not only from Synthetic B, but also from the four less preferred PEP's. Among the six preferred PEP's based on April data, the average rms difference is 0.07%. Differences between these and the two preferred August PEP's are larger, but we explained in our paper why we thought the April and August data were different. More

importantly, all the 14 adjustments improve upon the census by shifting population shares from areas where census-taking problems were low to areas where they were high. The fact that some of the adjustments, *e.g.* Synthetic B, make only small adjustments is no argument against making any adjustment.

F and N raise some additional questions, each of which we easily dispose of. First, Freedman and Navidi appear to disagree with our strategy of incorporating information from several sources. However, there was nothing about the sources of information that made combining them inconsistent or unusual. Since we start from the proposition that additional information is generally a useful thing, we do not find any merit in F and N's criticism on this point. Second, finding that the demographic method does not give a decomposition of the undercount that is geographically detailed, they set it aside as if it had no use. For us, the demographic method gives at least two important pieces of information. It gives a reliable estimate of the national undercount, and it also gives a powerful covariate: Blacks are undercounted more than Whites. Neither of these estimates should be taken to be without error, but they certainly give us confidence that each of the preferred PEP series coincide with these observations.

Finally, they found irrelevant our Table 6, which showed that omission rates, relative to rates of erroneous enumeration, were high in those areas with high undercounts. Turning to EKT's Table 5, we find it to be consistent not only with Table 6, but with the results of demographic analysis. This increases our confidence in the utility of the PEP. The argument is called "convergent validity," and is commonly made in the social sciences. It should also be noted that the series we do not take seriously because of the implausibility of their assumptions, Series 10-8, 14-8, 14-9, and 14-20, are less coincident with demographic analysis. We find:

1. Validation of the 8 preferred series, because their national undercount rates and the results in areas in which Blacks are concentrated are consistent with the demographic results, and
2. Evidence that Series 10-8, Freedman and Navidi's foil, is indeed an outlying series.

Can We Expect Improvements in Small Areas?

In our court testimony, we were concerned mainly to show that improvements could be expected for the 66 areas defined as PEP sampling areas. In a separate document, Tukey (1983) showed that if improvement was to be obtained in larger areas, then it could also be expected on average in its smaller components. Since then, both conceptual advances and empirical verifications (Ericksen *et al.* 1991, Appendix H; Wolter and Causey 1991) have been obtained.

Averaging and Sensitivity Analysis

Freedman and Navidi assert that "it is the spread in the PEP series that is interesting, not the average – because it is the spread that demonstrates the impact of applying different modeling assumptions to the same data (p. 11)." We differ from F and N in two ways. First, we believe that **both** the spread and the average are relevant, and we discussed each. Second, and more importantly, we used a different measure of the spread, the root mean squared error (rmse) instead of the range. F and N give little argument to support their choice of statistic. We prefer the rmse because it takes all the data into account, and the squared error feature gives extra weight to large errors. We found that "The root mean squared error among all 792 residuals is 0.59. In contrast, the root mean square of the 66 area effects is 1.60. The area effect is more than double the root mean square residual 47 of 66 times (EKT, p. 938)." We also showed that when we restricted attention to the "preferred eight," that the root mean squared residual was 0.33, and that the area effect was more than double the rmse 59 of 66 times.

We believe that F and N's use of the range in Table 6 and Figure 1 is wrong for another reason. Even among the preferred April estimates there is some difference among the national rates of undercount. If, as F and N say, we are concerned with shifts in shares of population, we should be concerned with deviations from the national average, as in our Tables 10 and 11. For example, in Florida the 2-20 estimate is 2.63% and for 3-8 it is 1.42%, for a range of 1.21%. Subtracting the national rates of 1.9 and 1.0 percent, the respective deviations are 0.73 and 0.42 percent, for a range of only 0.31%. Use of this statistic weakens the correlation displayed by Freedman and Navidi.

Assumptions

Freedman and Navidi argue that some of the assumptions underlying our regression model are "unverified" and "implausible." As we have already argued, both in the EKT paper and elsewhere (Ericksen 1986; Kadane 1986) we believe that they are both realistic and based on a body of knowledge that has been collected for decades. F and N assert that our model improves upon the synthetic estimates only if it uses additional information in a sensible way, bringing us right back to assumptions. We believe, despite F and N's assertions, that our assumptions are surely sensible, and indeed more realistic than the assumptions underlying a decision not to adjust.

At the same time, we believe that it is possible to make too much of the role of modeling in undercount estimation. For small areas, some type of modeling is surely needed. For the 66 areas our article was concerned about, the modeling did not usually make a lot of difference. For example, if we compare the mean residuals in our Table 11, which average residuals from the "preferred eight" estimates, with the corresponding mean residual of the eight sample estimates we find the following. For the 50 states, 46 of the residuals are within one percent of each other, and 48 are within one and one-half percent. The two remaining states, South Carolina and Tennessee, as we explained in EKT, appear to have sample estimates that are wrong, and the use of the regression model seems to provide a clear improvement. Turning to the 16 cities, five of the differences are in fact greater than two percent. For these, the sample sizes were smaller, and the weighted average is much closer to the regression estimate than to the original sample estimate. Although F and N would prefer us not to calculate the weighted average, we prefer to let the sample data play some role, perhaps small, to account for factors not necessarily included in the regression model. Either way, although we hold to our claim of their sensibility, we believe that the argument should be focused more on the quality of the PEP data than on the assumptions of our estimation model.

Does It Matter Which PEP Series is Used?

F and N hold to their position that there is no good reason to choose one PEP series over another. On the contrary, while it may be difficult to select a series from among the "preferred 8," there is good reason not to include Series 10-8 in this group. It is no solution simply to drop the movers from the analysis, as was done for Series 10-8, just because the August CPS had a problem identifying the April address of movers. As F and N themselves recognize, and as we learned from the PEP, movers had higher rates of omission and undercount. The problem with Series 10-8 is indicated in two additional ways. First, its national undercount rate, 0.3%, is well below the 1.4% estimated by demographic analysis. Second, the between-area variability is unrealistically too small, as we show in Table 5. The shift in shares created by Series 10-8 is similar to that of Schirm and Preston's Synthetic B which, while it improved over no adjustment, clearly did not go far enough. As a result, the between-area variability

among the 10-8 estimates for our 66 areas is too low. For example, assigning equal weights to each of the 66 areas as Freedman and Navidi appear to have done, the between-area variance for the Series 2-9 estimates is more than twice the corresponding between-area variance for the Series 10-8 estimates. Relative to the national average, the Series 10-8 estimates are too low in the high undercount areas and too high in the low undercount areas. It is little wonder, then, that the residuals from regression are a little bit smaller for Series 10-8 than for 2-9, and F and N's Table 7 has no real meaning.

Which Explanatory Variables Should Be Used

Freedman and Navidi believe that when Series 2-9 was the dependent variable in regression, we misapplied our own rules to select the independent variables. They argue that we should have added the percent living in poverty to our three selections – the minority percentage, the crime rate, and the percent conventional. This is because all four predictors in their equation (11) have coefficients which are more than twice their standard errors, and this equation has a smaller rms residual. They go on to assert that because the coefficient for the poverty variable was negative, we rejected the equation that use of our statistical criteria would otherwise obligate us to select. In other words, they assert that we let our subjective preconceptions overrule our statistical sense.

The problem with F and N's criticism is that they did not replicate our selection procedure correctly. As we explained in the article, and elsewhere (Ericksen and Kadane 1985; 1987, Section 6), "Our estimate of the undercount rate is a matrix-weighted average of a regression estimate and the initial sample estimates (EKT, p. 935)." The observations were weighted by the inverses of the standard errors of the initial sample estimates. This matters, because some states, like South Carolina, had aberrant sample estimates and large variances, and the sample sizes for the 16 cities were also smaller, causing the sample estimates to be less precise. Weighting the data by this procedure, for example, reduced the proportion of total weights assigned to the cities from 24 to 12 percent. When the poverty variable was added to our chosen three in a weighted regression, its coefficient was less than twice its standard error, and it was therefore excluded.

Freedman and Navidi also mistake a statistical decision for substantive motivation. On the contrary, had the poverty variable, with its negative coefficient, satisfied our statistical criteria, it would have added interesting and useful information to our estimates. In general, there are two types of areas with high rates of poverty, central cities with substantial minority populations and rural areas in states like Kentucky and West Virginia with small minority populations. Census errors are more likely to occur in either type of area than elsewhere, but the nature of the errors differ. In the cities, omission rates, as Table 6 in EKT demonstrates, were high, but in the rural areas, the rates of erroneous enumeration were high.

The effects of adding the poverty variable can be seen by subtracting F and N's equation (11) from equation (12), providing the following:

$$\text{difference in 2-9 fit} = 2.23 + .041 \text{ min} - .010 \text{ crime} + .001 \text{ conv} - .176 \text{ pov.}$$

In areas where the percents minority and living in poverty are both high, or both low, the difference may not be great. In areas with many minorities, but perhaps a slightly higher than average rate of poverty, the difference may be positive, but in areas with few minorities, but a high rate of poverty, the difference is negative. Of the 66 areas, the difference obtained from the above equation exceeded one percent only four times, and fell between 0.8 and one percent an additional six times. The ten most extreme areas are:

Area	Equation 12	Equation 11	Difference
		percent	
Maryland R	2.3	1.2	1.1
Houston	5.3	4.2	1.1
Washington, DC	8.1	7.2	0.9
Cleveland	4.3	5.1	-0.8
Arkansas	-0.3	0.5	-0.8
Mississippi	1.0	1.8	-0.8
South Dakota	0.4	1.3	-0.9
Kentucky	-1.3	-0.4	-0.9
Saint Louis	5.5	6.6	-1.1
Boston	3.4	4.9	-1.5

If we simply apply equation (11) to the 66 areas, with no averaging with the initial sample results, the shift in shares is as follows: Group 1, +0.36%; Group 2, +0.20%; Group 3, -0.56%. Substituting equation (12) we get: Group 1, +0.33%; Group 2, +0.21%; Group 3, -0.54%. While the difference between equations (11) and (12) is easily explained, and is consistent with our theory of census error, it really makes little difference to the final results.

Freedman and Navidi also return to the question of whether it was just as reasonable to use the percent urban as the crime rate. As we explained in EKT, use of the crime rate produced a lower rms residual and smaller standard errors than use of the percent urban. In their Tables 7 and 8, F and N appear to get different results. The discrepancy is explained by the same mistake noted above. By using unweighted data, they did not replicate our regression procedure, hence they got different results. Since their strategy gives greater importance to the cities, which had smaller sample sizes and therefore more uncertainty, it is not surprising that the percent urban becomes more important in their criticism.

Perhaps Freedman and Navidi think that our decision to weight the data by their estimated reliability is yet another arbitrary decision. Weighting seems obviously correct to us and is consistent with the strategy the Census Bureau followed in 1990. Where the observation seemed to be more reliable, we gave it greater weight. However, because they did not weight the data, much of F and N's analysis is simply different from ours, and their results in this article are not pertinent to what we did. This applies to their simulation study as well, both in this paper and in Freedman and Navidi (1986). Had they weighted the data, F and N may well have gotten different results. Even so, the fact that the variables selected for regression differ is not the real issue. The real issue is how much the actual estimates obtained from the different regression equations vary. The answer to that, as we have shown above, is that the undercounts do not differ substantially.

Final Comments

Perhaps the main point of the EKT paper is that within the range of reasonable PEP series, for any set of predictor variables that are well correlated with the undercount, results of undercount estimation are similar. In the end, the resulting undercount estimates are rather insensitive to changes either in the predictor variables or the choice of a PEP series. By a similar token, we do not give much weight to F and N's simulation results. The fact that different simulations adding random errors find different "best sets" of predictor variables does not tell us much, unless the distribution of the undercount turns out differently, which it does not.

In the absence of direct evaluation data, we carried out our sensitivity analysis, to see what the effects of various assumptions had on the estimates. In our view, substituting reasonable alternatives for the PEP series and undercount predictors made little difference. Moreover, the results followed a very reasonable pattern in light of the well-documented history of census-taking problems. In those areas where the Census Bureau had greater problems taking the census, the rates of omission, erroneous enumeration, and undercount were higher. In the end, we believe that the substantial and largely unchallenged evidence of serious census-taking errors combined with the consistency of estimates across choices of independent and dependent variables, and the agreement of the pattern of undercount with results of demographic analysis, provides ample reason to adjust.

Freedman and Navidi hold the adjustment data to a higher standard than unadjusted data. They take on faith, and contrary to decades of Census Bureau evidence, that the unadjusted data are accurate, and they do not seem to be concerned with an evident pattern of bias across areas. At the same time, and in the absence of any direct evidence, they assume large biases in the PEP data, when the Census Bureau studies do not demonstrate the existence of such biases (U.S. Bureau of the Census 1988, Section 6F) In other words, they do not seem to place the unadjusted and adjusted data at the same starting point when making their analysis. In doing so, F and N are able to throw out "possible problems" as if they were real ones and to neglect real problems with the unadjusted census as if they did not exist. They reject adjustment on this basis alone.

ADDITIONAL REFERENCES

- ERICKSEN, E.P. (1983). Affidavit presented to U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).
- ERICKSEN, E.P. (1986). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. *Statistical Science*, 1, 18-21.
- ERICKSEN, E.P., and KADANE, J.B. (1986). Using administrative lists to estimate census omissions. *Journal of Official Statistics*, 2, 397-414.
- ERICKSEN, E.P., and KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. In *Small Area Statistics: An International Symposium*. (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh). New York: John Wiley & Sons.
- FREEDMAN, D.A. (1984). Testimony given in U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).
- KADANE, J.B. (1986). Comment on, Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi, *Statistical Science*, 1, 12-17.
- TUKEY, J.W. (1983). Affidavit presented to U.S. District Court, Southern District of New York, in *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).
- U.S. BUREAU OF THE CENSUS (1985). The Coverage of Housing in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E1. Washington, DC: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1987). Programs to Improve Coverage in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E3. Washington, DC: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1988). The Coverage of Population in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E4. Washington, DC: U.S. Government Printing Office.

RESPONSE FROM THE AUTHORS

1. Introduction

After some general remarks, we respond to each of the discussants' main points. There is some overlap among their arguments; we try to deal with each point only once. Like the other participants, we have learned something over the years – and from the present exchange – but have not changed our opinions on the central questions. One issue cannot be in dispute: Editor M.P. Singh deserves thanks from all sides.

2. A brief Outline of Adjustment

There is a proposal to adjust the census using capture-recapture techniques. A person is “captured” if they are counted in the census; “recapture” is in a special sample survey done after the census. In 1980, this survey was called PEP, or Post Enumeration Program. In 1990, the terminology shifted to PES, or Post Enumeration Survey.

These surveys measure the rate at which people are missed from the census (“gross omissions”), as well as the rate at which people are counted in error (“erroneous enumerations”). Erroneous enumerations include babies born just after census day, people counted at the wrong address, *etc.* To a first approximation, the net undercount is estimated as the difference:

$$\text{gross omissions} - \text{erroneous enumerations}.$$

There is a significant additional complication. In 1980, sampling error was a large-enough problem (according to many observers) so that estimates from the survey could not be used directly. Instead, in EKT's terminology, “sample estimates” from the PEP had to be run through a smoothing model to get “composite estimates.” In 1990, the terminology is different: “raw adjustment factors” from the PES are modeled to get “smoothed adjustment factors.” But the problem of sampling error is even more salient. For more details, see Freedman (1991), U.S. Department of Commerce (1991a, pp. 4.2-4.18), or Wolter (1991).

3. The Census is Bad so the Alternatives must be Better

Many discussants make an argument which, baldly summarized, comes down to this: the census is bad; the PES must be better; therefore, we should adjust. This is a confusion: it treats the census and the PES as alternatives. However, you cannot choose the survey instead of the census; at most, you can try to use the PES to correct flaws in the census. The question, then, is not whether the survey is better, but whether it is good enough for its intended use.

The Secretary of Commerce framed the issue as follows:

“I concede the census' imperfections, but the critical inquiry . . . is not how flawed the census is, but whether the PES can fix it . . . [W]hile identifying flaws in the census is important for planning the next one, it simply begs the question . . . Is there convincing evidence showing that the adjustment is more accurate than the enumeration? [U.S. Department of Commerce 1991a, p. 2.13].”

4. Fienberg

Fienberg defines the central issue as follows:

“At issue is both the accuracy of the census and the adjustment process. And, it is the substantial differential undercount, *i.e.*, the difference between the undercount for Blacks and the undercount for non-Blacks and between Hispanic and non-Hispanic, that is important when we come to assess census accuracy. This is because census figures are typically used to divide resources among groups in the population, resources such as seats in the U.S. House of Representatives; seats in state legislatures; federal funds; and so on. [p. 25, emphasis omitted].”

We think this is misleading. The argument is about shares: more specifically, the accuracy of shares computed from adjusted figures and from the census. But the shares that matter are for geographical areas – states, cities, counties, and so forth. The total share of blacks or hispanics in the U.S. population, at the national level, matters much less. Seats in Congress are allocated to states, and within states to geographical areas. They are not distributed to national racial or ethnic groups. Similarly, tax moneys go to some 39,000 local governments, defined by area not race or ethnicity. The crucial issue is whether adjustment improves the accuracy of population shares for geographical areas rather than groups.

Fienberg is misleading at other points as well. We give two examples.

- (i) Fienberg (p. 27). “[Freedman and Navidi] focus on the variation amongst the full set of 12 alternatives, some of which to me are implausible given the assumptions that they rely upon.” But we did study variation among EKT’s preferred series rather than the full set: see pp. 7-8 and 13-14. We did this not because we agreed with EKT’s choices, but to make irrelevant Fienberg’s kind of argument. That didn’t stop him.
- (ii) Fienberg (p. 27). “I read the report by Ylvisaker (1991) who reexamined data from a trial census in Los Angeles in preparation for 1990, but I could not find the evidence that Freedman and Navidi state is supportive of their claim that smoothing increases variability.” Ylvisaker did a bootstrap experiment using data from Los Angeles, where there was a test census and a test post enumeration survey in 1986. At the tract level, bootstrap SEs for the smoothed estimates are generally larger than the SEs for the raw estimates. (See Ylvisaker’s Table 3; smoothing reduced the SEs in 19/61 tracts, increased the SEs in 26/61, and the remaining 16/61 were ties; at the block level the effects go the other way but are small in either case.) For the whole site, the comparison is as follows (Ylvisaker p. 7):

SE for smoothed estimate = 0.75.

SE for raw estimate = 0.68.

As we said (p. 17), “smoothing may actually increase sampling error.”

Nothing is Perfect, and don’t Let the Best be the Enemy of the Good

Fienberg says (p. 27),

“A familiar theme in various writings by one of the present authors is the problems that arise when assumptions are not satisfied. Here again the authors pursue this theme with respect to the linear equation used for smoothing. They appear to argue that either all assumptions are perfectly justified or ‘all bets are off.’ Nothing could be further from the truth.”

Alas, our position is more complicated than that. We think the census is imperfect, but good. We think the smoothing models are quite questionable, and the arguments to defend them are bad. Proponents of adjustment have an obligation to state their assumptions and produce data to validate them. The models don't have to hold perfectly, but departures from assumptions and their impacts need to be studied. Otherwise, the algorithms have no justification except familiarity.

5. The Burden of Proof

As the exchange with Fienberg indicates, modelers are reluctant to accept the burden of proof. Once they make an assumption, it is taken as truth unless it can be disproved. Even then, they may view the assumption as useful until it can be replaced by some other assumption.

Language is used in a specialized way. An assumption is "reasonable" if the modelers think it is reasonable. If questioned, they introspect again. The introspection confirms the original conclusion; after all, the assumptions are by now familiar parts of the technical literature. The modelers become indignant at those who do not share the faith. If all "reasonable" options favor adjustment, arguments on the other side must be "unreasonable."

So far as they are reported, the modelers' thought experiments do not seem especially rigorous; and the pro-adjustment argument can be peculiarly non-empirical. Illustrations follow. One axiom in the Ericksen-Kadane smoothing model is independence. See equations (1-6) in our paper. Independence drives the variance calculations, because small correlations can have a big cumulative impact. Variances determine whether smoothing is a help or hindrance. The independence assumption matters.

As far as we can see, the adjusters' main arguments for independence are the following:

- (i) The errors are not perfectly correlated. (We adapt to present context an argument by Madansky 1986, p. 29.).
- (ii) (a) "The 1980 census was administered by more than 400 district offices, an average of eight per state. (b) To our knowledge no one has suggested that there actually was an April snowstorm or any other event that affected the census in neighboring states. (c) When we correlated PEP estimates for cities with the corresponding estimates in their states (e.g., Detroit with the remainder of Michigan), we found no evidence of a correlation." [EKT, p. 931; we responded to (b) by noting the eruption of Mt. St. Helens.]
- (iii) "Surely they don't expect anyone to believe the argument that the eruption of Mt. St. Helens interfered with census taking in a serious way . . ." [Fienberg p. 27.]

In fact, what we expected from the modelers was serious argument about the validity of assumptions, rather than intuitions about possible sources of dependence like snowstorms and volcanoes. Over time, the force of that expectation has dwindled. Real empirical evidence is hard to get, on both sides. Their mainstay is the rhetoric: Nothing is perfect, so anything goes. That is the adjusters' standard for the models. On the other hand, the census is required to be right to within a few percentage points – where "right" is defined by the models.

6. Fellegi

We agree with many of Fellegi's points. In the U.S., for instance, small-area income data help determine funding allocations. These data have weaknesses of their own, not addressed by census adjustment. Likewise, there are substantial shifts in the population between censuses. Better income data, or a mid-decade census, might be more useful than any adjustments to the decennial census.

There is one point we would like Fellegi to consider. A decision to adjust the census, whether in the U.S. or in Canada, has major organizational costs: it encourages the replacement of data collection by modeling.

“In sum, real data (with real flaws) would be replaced by complicated and poorly tested mathematical models of data. We do not see that as progress.” [Beran *et al.* 1988.]

7. The PEP Series

In 1980, there was a substantial amount of missing data in the surveys used to assess census error. Different ways of filling in the missing data lead to different estimated undercount rates. In the end, the Bureau had a dozen different PEP series: each provides estimated undercount rates for 66 geographic areas (central cities, states apart from their central cities, whole states). A series is identified by a pair of numbers, *e.g.*, PEP 2-9 or PEP 10-8. For more details, and arguments about the merits of the various series, see FN p. 4, EKT p. 929, Fay *et al.* 1988, p. 63.

8. Cressie

Cressie agrees (p. 32) that in 1980, “data and methods were inadequate for an accurate adjustment of the whole country.” Of course, many of the arguments are relevant to the 1990 decision, and on those, Cressie’s opinion may differ from ours. This is not the place for an extended discussion of 1990, but we can respond to some of his points, at least in outline.

Demographic Analysis

Cressie – like other discussants – relies on estimates from demographic analysis, a technique that uses administrative records (birth certificates, death certificates, *etc.*) to make an independent estimate of the total population. For details, see Fay *et al.* (1988).

How good is demographic analysis? It may be surprising to some, but government statistical agencies keep changing their minds about the past. The estimated GNP for a year in the past – 1985, for example – depends on the year in which the estimate is made. The numbers keep on changing, and the revisions give some clues about the reliability of the initial data.

Table A gives a brief history of revisions to demographic analysis for the 1980 census. As will be seen, the numbers are far from stable. The difference between estimates made in 1984 and 1988 may reflect new understanding about the role of illegal immigration. The change from 1988 to 1991 may reflect the impact of adjustments to earlier adjustments intended to correct for under-registration of births in the period 1935-1960. Apparently, these were over-adjustments, which may now have been fixed.

Table A
A short history of revisions to demographic analysis of the 1980 census:
Estimated undercounts by date of estimate

	1984	1988	1991
All races	0.5	1.4	1.2
Blacks	5.3	5.9	4.5
Non-Blacks	−0.2	0.7	0.8
Differential	5.5	5.2	3.7

Source: Col. 1. Cressie, citing Passel and Robinson (1984); the figure for all races is derived.
Col. 2. Fay *et al.* (1988, p. 95, series DA-2).
Col. 3. U.S. Department of Commerce (1991c, Table 3).

Demographers can use data from administrative records to estimate the population of the U.S., and they seem to get it right to within a percentage point or two – a remarkable achievement. However, it seems unlikely that the errors are much less than a percentage point. If so, demographic analysis may not be reliable enough for adjusting the census.

Modeling

Cressie says (p. 33),

“The important concept to maintain is that true undercount in regions is unknown and the ignorance is quantified into a probability model. The goal is not estimation of the coefficients β but prediction of the undercount. With an error term that does not have to be independent and identically distributed, this prediction is insensitive to misspecification ... [emphasis omitted].”

We disagree. A model for one investigator’s ignorance is no basis for public policy. And results must depend strongly on specifications. To illustrate, we note some of the assumptions in the model developed by Cressie (1988). Equations (2.7) and (2.10) in that paper effectively rule out nonsampling error in the PES, as well as systematic variation in undercount rates across geographical areas; and no correlations appear. Why? Equation (2.10) specifies a sampling variance which avoids the internal inconsistencies in the Ericksen-Kadane model (Cressie 1988, p. 193). However, logical consistency does not imply empirical truth. Where does the real sampling design come in? Finally, why should we use Cressie’s loss function (2.15)? Until Cressie answers these questions, and others like them, his model outputs have no claim to be taken seriously.

When Cressie gets down to cases, he is computing estimated risks (expected losses). See his equations (2.28-2.31). That means he has to compute variances. Variances are extremely sensitive to assumptions, as Cressie knows:

Needless to say, these results rely on the correctness of the assumed model. [p. 193].

An elementary illustration may help. Suppose $\epsilon_1, \dots, \epsilon_{66}$ are exchangeable, with mean 0, variance σ^2 , and pairwise correlation ρ . Now

$$\binom{66}{2} = 2145.$$

Therefore,

$$\text{var}\{\epsilon_1 + \dots + \epsilon_{66}\} = (66 + 2145\rho)\sigma^2.$$

In this game, a correlation of, say, 0.05 makes a huge difference. And correlations that small would be quite hard to detect empirically. Cressie doesn’t try. (With 16 data points, even a correlation of 0.5 might be hard to estimate, so EKT’s test #3 on p. 931 cannot have much power.)

The example may seem artificial. However, sampling error was a major obstacle to adjusting the 1990 census on the basis of the PES, even at the state level. Indeed, published data show that for a clear majority of states, the population shares from adjustment would be within two standard errors of the census shares (U.S. Department of Commerce 1991b). Such adjustments could result entirely from sampling error in the PES. (“Loss function analysis” might be the adjusters’ response, and we discuss that briefly when answering Hartigan.)

The standard errors, like the estimated adjustments, are outputs from a smoothing model akin to the EK model. Bootstrap experiments reported in Fay (1992) show that these standard

errors are too small by a factor of 2 or so. (Fay gives the range 1.4 to 2.2, with a preferred multiplier of 1.7.) When it comes to computing variances, assumptions make all the difference.

PEP 3-8

On p. 32, Cressie agrees that the 1980 adjustment data were not strong enough to use. By p. 33, he wants to adjust using his model and PEP 3-8 (see Section 7 above). He seems to have assumed away all the problems created by non-sampling error, missing correlations, and so forth. If so, his calculations are unrelated to the policy questions.

The Quality of the PES

Cressie says (p. 33) that the PES was “well designed, well implemented, and quality assured.” So it was, relative to a typical market research survey, or perhaps even relative to other Census Bureau surveys. However, to fix a small error in the census, you need a sample survey which makes much smaller errors. And we do not believe the PES meets that standard. For example, the PES estimated a national undercount rate of 2.1%. Between 1/3 and 2/3 of that 2.1% can be attributed to non-sampling error in the PES. See Mulry (1991, Table 15) and Bryant (1992). The PES seems to be fatally flawed. We return to this topic in answering Hartigan, below.

Conclusion

Cressie’s main point seems to be this (p. 33):

“To solve a problem as hard as adjustment for undercount, the common goal needs to be recognized. From there, debate should center around differences on how that goal might be reached. If Freedman and Navidi’s position is that the goal is impossible to reach (which is what they seem to have implied over the years), then it should be stated.”

Let us be clear. In our opinion, PEP could not solve the problem in 1980, and the PES cannot solve the problem in 1990. Nor are we optimistic about the year 2000, whatever acronym may be in use then. If you can’t count them, you shouldn’t make them up afterwards by running capture-recapture data through smoothing models.

9. Passel (1987)

Many of the discussants defend synthetic adjustment, some very strongly. Few of them are much taken with our counter-example (Table 3). However, Passel (1987) used 1980 census data to show that synthetic adjustment was unlikely to improve accuracy. His work was summarized in the Appendix to our paper. No discussant responds to his argument.

10. Schirm and Preston

The Counter-example

SP (1987, p. 966) make a claim about synthetic adjustment:

“Our finding is that synthetic adjustment will always move the estimated ratio of a state’s population to the national population closer to the true ratio if (a) the state’s black undercount is closer to the national black undercount than it is to the national undercount for both races combined and (b) the state’s white undercount is closer to the national white undercount than it is to the national undercount for both races combined.”

Our counter-example (Table 3) showed this result to be wrong. They should concede the point.

Under some conditions, and by some criteria, synthetic adjustment is doubtless a good thing to do; see, *e.g.*, (15) in our paper. The result in SP's (1987) appendix is correct but not illuminating: the inequality they assume in equation (A.2) on p. 976 is exactly the inequality on absolute error they seek to prove, up to multiplication by a scale factor.

The Simulations

S and P prefer a strict definition of the synthetic assumption – “there is no variation at all” in undercount rates within race across geography. They say (p. 37) that they did not construct true populations on the basis of the synthetic assumption, and their definition of truth did not favor synthetic adjustment.

We adopt their terminology for a moment. They constructed the true populations from the synthetic assumption plus random error. Indeed, the simulations hold the census counts fixed, and randomize the true population. The true population of racial group j in state i is assumed to equal the corresponding census count, multiplied by a random adjustment factor u_{ij} . See SP (1987) equation (1) on p. 967. This adjustment factor is drawn at random from a distribution which depends – by assumption – on the racial group but not the state. See SP (1987) equation (2) on p. 967.

The simulations assume away systematic variation in undercount rates within race across geography. On the other hand, synthetic adjustment assumes that the structure of undercounts is determined by race not geography. That was our point on p. 10, and it is right.

Indeed, S and P concede (p. 37).

“We considered cases of extreme, albeit nonsystematic, interstate variation in undercounts by race”

The “albeit nonsystematic” is their concession; the “extreme” must be the defense.

The *a fortiori* Argument

S and P say their simulations were conservative; the real pattern of variation in undercount rates across areas would favor synthetic adjustment even more strongly than the assumptions they made. (See *e.g.* p. 38). SP (1987) had *a priori* arguments to that effect. Passel (1987) shows, among other things, that such arguments do not prove much about 1980; see the Appendix to our paper. SP (1987, p. 977) make some empirical arguments, using data that are “seriously flawed, based on heroic assumptions”; S and P's language (p. 38), not ours. Further discussion seems unnecessary.

On p. 38, S and P introduce new analysis based on PEP to justify the parameters in the simulations. In present context, that is quite a move: EKT want us to believe the PEP series because they are like the synthetics, while S and P want us to believe the synthetics because simulations are like PEP.

Before we accept either, we want some evidence. Circular reasoning is not persuasive.

11. Hartigan

Synthetic Adjustment

Hartigan rejects Schirm and Preston, but argues strongly in favor of synthetic adjustment (p. 45). He says,

“What about the following analytic argument? Suppose the national undercounts are correctly estimated, but the undercounts differ over states . . . National undercount rates of 5% and 1% are supported by historical data from the Bureau, both by demographic analysis and post enumeration surveys . . . I will ignore variations in the non-minority undercount between States . . .”

Unless we are much mistaken, these analytic arguments are too far from the facts to be relevant. Hartigan’s basic assumption is that the national undercount rates are known. That assumption is wrong: we doubt that the rates can be reliably estimated, either by demographic analysis or the PES, to within a factor of 2. See our discussion of Cressie, above. Furthermore, Hartigan ignores variations in non-minority undercount rates across states. Such variation has to matter: for example, a 1% undercount among 9 million people in a state has almost twice the impact of a 5% undercount among 1 million.

Modeling

Hartigan says

“I would suspect that the assumptions of the regression can not be easily defended, but that the results of the regression are reasonable, except perhaps in producing lower standard errors that are justified by the probable lack of independence . . . Reduction in sampling variance by regression-based smoothing procedures is not likely to make much difference to estimates in large localities such as States . . . A healthy skepticism about any resulting ‘standard errors’ or ‘confidence intervals’ is justified. [p. 48 emphasis omitted].”

For 1980, the choice of variables makes a lot of difference to the adjustments for small areas. See FN p. 9. For 1990, the “raw” adjustment factors (computed directly from the sample without regression) have such large sampling errors as to be unusable, even at the state level. So the adjusters need to smooth. But the choice of smoothing models makes quite a difference to the results. See the Secretary’s Decision (U.S. Department of Commerce 1991a, pp. 2.46-2.55) and consider the numbers in the Press Release (U.S. Department of Commerce 1991b).

Furthermore, the argument for adjusting rides on a “loss function analysis,” which uses variances computed from the smoothing model to make unbiased estimates of risk. The model is known to be too optimistic about its variances, perhaps by a factor of 5; see FN p. 10, Ylvisaker (1991), our main paper Section 7.3, and Fay (1992). If “healthy skepticism” is applied to the loss functions, we see no arguments left on the table for the efficacy of proposed adjustments.

We expect to discuss the Bureau’s loss function analysis in another paper. Hartigan does his own calculations on p. 49; again, they are too far removed from the data to carry much weight. In any event, readers can look at the Bureau’s analysis (Mulry 1991; Woltman *et al.* 1991) before buying any conclusions.

The Third Pope

“The bureau has produced a number of estimated undercounts, with margins of error, in the various states. I use the ‘selected PES method’ (called PES from now on) . . . Now there are two popes, the enumeration and the PES figures. Which is correct? Well, you need a third pope, an infallible one . . . [p. 49].”

Hartigan is on to something important here. The Bureau’s “third pope” consists of the loss function analysis discussed above, and a “total error model” (Mulry 1991). These seem highly fallible: the loss function analysis because it depends on variances computed from the smoothing

model, and the total error model because it depends on results from the Evaluation Followup to measure non-sampling error. (Furthermore, the two models interact in crucial ways, but that is a topic for another day.)

The adjusters are trying to fix an undercount of maybe 2%. To do that, they need to control non-sampling error in the PES to well below 1%. They say they did it, on the basis of data from yet another sample survey – the Evaluation Followup. If they are measuring non-sampling errors in the PES to within a fraction of 1%, the errors in the Evaluation Followup have got to be an order of magnitude smaller. They must be kidding.

The Five Questions

Hartigan concludes with five questions, and we will answer two. (The first is edited slightly, for clarity.)

(i) “Do Freedman and Navidi agree with these estimates of 17 million omissions and 13 million erroneous enumerations?” We accept the numbers as rough estimates, subject to large and unknown biases as well as large and unknown standard errors. The difference of $17 - 13 = 4$ million may be off by a factor of 2 or more. Estimating a small number by taking the difference of two large numbers is a time-honored recipe for trouble.

Furthermore, a crucial issue is where to put the 4 ± 2 million people. Fienberg doesn’t like the foothills of South Dakota. That narrows the options to 6.5 million blocks spread over 39,000 jurisdictions. The PES gave us data on 0.2 of 1% of the blocks, and perhaps 10% of the jurisdictions. Great theater compels the audience to suspend disbelief. Adjustment does not reach that level.

(ii) “If the PES is not good enough, how should the follow-up survey be designed so that it could be used to adjust the census?” The answer is a question of our own: What on earth makes him think it can be done at all?

12. Speed

Adjustment depends on models and assumptions for which there is no empirical proof. That is Speed’s message, and we agree.

To adjust the 1990 census, the population is divided into 1,392 “post strata,” or demographic groups. One example is post stratum 90302112, male hispanic renters age 10-19 in cities in the Pacific Division. The adjustment depends on the “homogeneity assumption,” that undercount rates are more or less constant with each post stratum across geographical areas. See Freedman (1991) or (U.S. Department of Commerce 1991a, pp. 2.37-2.45, pp. 4.16-4.18).

This assumption is hardly an obvious truth. The Bureau did some work to test it (Kim 1991). However, that study seems to have been quite poorly designed, and in any case gives rather mixed results. The theory of adjustment is particularly shaky when it comes to small areas.

13. Ericksen and Kadane

The Role of Assumptions

We say that “success of any of EKT’s proposed adjustments rides on unverified and implausible assumptions.” EK answer (p. 52) that their “assumptions are realistic and verified by decades of census-taking knowledge, as we will argue below.”

The argument they have in mind seems to be on p. 55:

“Freedman and Navidi argue that some of the assumptions underlying our regression model are ‘unverified’ and ‘implausible.’ As we have already argued, both in the EKT paper and elsewhere (Ericksen 1986; Kadane 1986) we believe that they are both realistic and based on a body of knowledge that has been collected for decades.”

We reviewed the EKT paper, as well as (Ericksen 1986) and (Kadane 1986). We found no empirical evidence to substantiate the assumptions, or to quantify failures (*e.g.*, to determine the real sizes of the correlations assumed to be 0), or to determine the impact of failures on model output. Instead, EK rely on arguments from convenience (a good model is “simple and tractable” and “permits smoothing,” Kadane 1986 p. 13). They also have their own variation on nothing-is-perfect rhetoric:

“... in applications, only a very naive user would believe in the literal truth of the assumptions. Thus in my view, when I state and use an assumption, I mean that I think something like this is true, but surely I do not mean that exactly this is true ... (Kadane 1986, p. 14).”

What makes EK think that “something like” their model is true? Convenience and nothing-is-perfect, even in combination, do not validate assumptions or quantify the impact of failures.

Opening another front, EK tax us with having our own unverified assumptions. Our guilt on this score would hardly imply their innocence; but we deny the charges, or at least most of them. Three examples give the flavor of our “unverified assumptions” (p. 53).

- (i) A small undercount is thought to remain in the census.
- (ii) Minority persons living in central cities are likely to behave differently from those in suburbs.
- (iii) The undercount in conventional areas was relatively high.

Point (i) still seems to be right. If EK will concede that it is wrong, we can all save a lot of courtroom time and journal pages. Point (ii) is obvious to anyone who has spent a few days in a big city in the U.S., but if data are needed, see Freedman *et al.* (1991), which also reviews some literature on this topic. On point (iii), we should have said “estimated undercount.” Touché.

Evidence for Assumptions

One example is enough. EK say (p. 52) there is

“evidence of greater census-taking problems in some areas than others, and ... PEP-measured omission rates and undercounts are higher in those areas with lower mailback rates, higher rates of missing data, and greater problems maintaining the specified long-form sampling rate on the census.”

(For a brief review of the PEP series, see Sections 2 and 7 above.) At most, EK are proving that the PEP data have some relationship to undercount rates, and that we never denied. However, not all relationships can be summarized in a regression model. To get the model going, EK (p. 54) say only “there was nothing about the sources of information that made combining them inconsistent or unusual.” This is astonishingly weak, because the tests are only the following: (i) the model should have no internal contradictions; (ii) somebody else should already have done something similar.

Adjusting Small Areas

Earlier, EKT seemed to concede that they could not adjust small areas (p. 943, also see Section 4 of our main article). EK now withdraw the concession (p. 54), citing work by Tukey and Wolter and Causey. That work was reviewed in the Appendix to our paper. We do not find it convincing, and explained why. EK do not respond to our arguments.

EKT's Table 5

EK say that our argument "goes too far." However, their Table 5 is supposed to show that their preferred PEP series are in general agreement with synthetic estimates. Such agreement would demonstrate the value of PEP only if synthetic estimates were known to be accurate. That premise is doubtful, as discussed before.

Furthermore, on the scale EKT chose, we found remarkable disagreement among their preferred PEP series. EK's response: they previously restricted attention to 8 of the 12 PEP series, but now want to eliminate two more (from August). That is not good: among other reasons, the extreme difference noted in our equation (8) occurs with April series making the final cut. Next, EK average across their most preferred series. Averaging results from a sensitivity analysis to reduce variation is a peculiar idea, as discussed in Section 5 of our paper. We return to the point, below.

EK go on to say (pp. 53-54):

"More importantly, all the 14 adjustments [the 12 PEP series and the two synthetic adjustments] improve upon the census by shifting population shares from areas where census-taking problems were low to areas where they were high."

This is euphemistic. As Table 5 in EKT makes clear (and see EKT p. 927, EK p. 53), the areas where census-taking problems were high are the areas with a high concentration of minority persons. However, as we explained in responding to Fienberg, legislative seats and tax moneys are allocated to geographical areas, not racial or ethnic groups. The key issue is whether adjustment would improve the accuracy of population shares for small geographical areas – states, cities, counties. EKT's Table 5 is about broad groupings of cities and states. Such aggregates seem artificial.

Which PEP Series?

EK try yet one more time (p. 54) to justify their preference for 8 out of the 12 contending PEP series; they particularly seek to eliminate our dreaded foil, series 10-8. The main argument is concordance with demographic analysis at the national level. EK also claim that "the results in areas in which Blacks are concentrated are consistent with the demographic results." This must be a slip in the prose, since demographic analysis gives no results below the national level.

EK indicate (p. 54) that concordance matters, because demographic analysis "gives a reliable estimate of the national undercount." Demographic analysis probably is more reliable than PEP. But it has real problems of its own: see our discussion of Cressie, above. Concordance is a weak argument.

Furthermore, any agreement between PEP and demographic analysis at the aggregate level masks substantial differences in detail, as Jeff Passel showed in court. The arguments have been reviewed before, but we try again. PEP 2-9 is the most preferred of EKT's preferred PEP series; Table B compares PEP 2-9 to demographic analysis. PEP 2-9 is a bit low on black males, 100% too high on black females, 33% too low on white males, and too high on white females (0.5 of 1% vs. 0). The agreement has evaporated.

Table B
Comparing two ways of estimating undercount rates in the 1980 census:
Demographic analysis (DA) and PEP 2-9.

	DA	PEP 2-9
Black Males	8.8	8.1
Black Females	3.1	6.4
White Males	1.5	1.0
White Females	0.0	0.5

Source: Fay *et al.*, 1988, Appendix D.
Note: Demographic analysis is based on the series DA-2; “white” includes “other races.”

Moreover, EK’s defense is not totally consistent. For instance, compare p. 54 with p. 55. On p. 54, national undercount rates are important, on p. 55, variation in national undercount rates is unimportant. And their position of the moment – averaged across pages – is inconsistent with that taken by the National Academy of Science Panel, where prominent participants were Steve Fienberg and Jay Kadane:

“There are a number of reasons, both *a priori* and *a posteriori*, supporting the various individual [PEP series] from this list of 12 For example, estimate 10-8 reduces the problem for movers when using the August P-sample These points among others are detailed in Bailer [’s affidavit in *Cuomo v. Baldrige*].”

“The use of these 12 estimates produced very different estimates of undercoverage for national demographic groups Some analysts have suggested that the number of acceptable estimates should be narrowed considerably. For example, Ericksen . . . would discard all but the 2-8, 2-9, 3-8, and 3-9 estimates as either based on August data, which had a higher rate of cases with unresolved match status, or as making use of extreme assumptions in the adjustments for missing data. However, even within this restricted set, the national undercount rate ranges from 0.8 to 1.4 percent. [Cohen and Citro 1985, pp. 147-148.]”

In short, even among EK’s most preferred series, different imputation models give different results. Nor is there good reason to discriminate against our foil 10-8.

Likewise, EK say (p. 55) that “Schirm and Preston’s Synthetic B . . . , while it improved over no adjustment, clearly did not go far enough.” This contradicts previous positions taken by the panel, albeit tentatively; see (Cohen and Citro 1985, p. 287; our paper, p. 8).

Averaging and Sensitivity Analysis

EK invite us (p. 54) to replace the various PEP series by the average, and to consider rms deviations from average. However, the point of the 12 different imputation schemes was to measure the impact of modeling. For that purpose, the range is the right statistic: two randomly selected imputation models may give similar results, yet a third may be quite different. In the end, EK want to do a sensitivity analysis, but downplay any model that is different from the other ones.

EK propose again to subtract each series’ estimated national undercount rate from its estimates for the 66 study areas. EK are tacitly assuming – with no basis – that one imputation model holds for the whole country. Our analysis takes the view that data may be missing for different reasons in different parts of the country (Section 7.1).

EK give new reasons (p. 55) for rejecting PEP 10-8. The argument comes down to this: if their preferred series are right, our foil 10-8 is wrong. Just so. Conversely, if 10-8 is right, their series are wrong. In other words, it matters which PEP series is used. When you are estimating small undercount rates, 8 percentage points of missing data make a difference. No statistical manipulation can change that awkward fact.

Replication

EK say,

“The problem with FN’s criticism is that they did not replicate our selection procedure correctly. As we explained in the article, and elsewhere (Ericksen and Kadane 1985; 1987, Section 6),

- (i) The observations were weighted by the inverses of the standard errors of the initial sample estimates.
- (ii) ‘Our estimate of the undercount rate is a matrix weighted average of a regression estimate and the initial sample estimates.’ [p. 56, order of points interchanged from original.]”

Regrettably, EK are confounding two issues: (i) how you select variables, and (ii) what you do after selecting them. With respect to point (ii), EK are using the Lindley-Smith hierarchical Bayesian regression model. There is only one wrinkle: the parameter σ^2 is unknown and must be estimated; see FN pp. 5 and 11.

Once the variables are selected and σ^2 is estimated, there is no ambiguity about EK’s estimator. See Ericksen-Kadane (1985) equation (3), FN equation (9), or the notes to Table 8 in our main paper. Indeed, we were able to replicate their numbers in court; the judge even complimented us on the accuracy of the Berkeley computers. We illustrate the point again, with data in EKT. Their composite estimator based on PEP 2-9 can be extracted from Table 10. The lead example is St Louis, and their value for the estimator is

$$1.24 + 0.66 + 4.16 + 1.09 = 7.13.$$

Our value is 7.12. (The largest discrepancy we found was for Dallas: 6.22 vs. 6.18.) Given the variables, we can do the rest.

Most of the discussion in FN, and in the present paper, depends on what you do after selecting the variables, and is immune from EK’s criticism of incorrect replication. In particular, despite EK having singled it out by number, our Table 8 is fine. It has nothing to do with the algorithm for variable selection, and we stand by it.

The situation is otherwise with our equations (11-12), Table 7, and Tables 9-10, corresponding to Tables 5 and 6 in FN. Those calculations really do depend on the variable selection algorithm, and we discuss the implications after a few remarks to provide context.

In 1986, EK criticized our simulations, but not on present grounds: the simulations started from the infamous series 10-8. The issue of OLS vs. GLS was not mentioned. In 1989, EKT criticized the simulations again, for yet another set of reasons: (i) we restricted attention to models with 3 variables, and (ii) we did not require the coefficients to be significant.

They raise the issue of OLS vs. GLS now, for the first time. In response, we redo our calculations once more, using GLS with observations “weighted by the inverses of the standard errors of the initial sample estimates”; coefficients must be significant, but negative values are permitted. We report first on equations (11) and (12); *t*-statistics are shown in parentheses.

$$\begin{aligned} \text{(OLS 11)} \quad \text{PEP 2-9} &= -2.23 + .079 \text{ min} + .036 \text{ crime} + .028 \text{ conv} + \text{residual} \\ &\quad (-4.0) \quad (5.4) \quad (3.6) \quad (3.5) \\ \text{rms residual} &= 1.53. \end{aligned}$$

$$\begin{aligned} \text{(OLS 12)} \quad \text{PEP 2-9} &= .120 \text{ min} + .026 \text{ crime} + .029 \text{ conv} - .176 \text{ pov} + \text{residual} \\ &\quad (7.6) \quad (3.4) \quad (3.8) \quad (-4.4) \\ \text{rms residual} &= 1.49. \end{aligned}$$

$$\begin{aligned} \text{(GLS 11)} \quad \text{PEP 2-9} &= -3.37 + .054 \text{ min} + .061 \text{ crime} + .026 \text{ conv} + \text{residual} \\ &\quad (-6.0) \quad (3.6) \quad (5.4) \quad (5.0) \\ \text{rms residual} &= 1.60. \end{aligned}$$

$$\begin{aligned} \text{(GLS 12)} \quad \text{PEP 2-9} &= .118 \text{ min} + .030 \text{ crime} + .031 \text{ conv} - .217 \text{ pov} + \text{residual} \\ &\quad (7.3) \quad (4.1) \quad (5.2) \quad (-5.4) \\ \text{rms residual} &= 1.53. \end{aligned}$$

Min is the percentage of minorities; crime, the crime rate; conv, the percentage who were conventionally enumerated; pov, the percentage below the poverty line.

As will be seen, the weights make little qualitative difference (although the difference in *t*-statistics is noticeable). Under either regime, pov is quite significant. And the equation involving pov is superior, for it has smaller residuals.

The poorer an area is, the *less* its undercount will be. That is what equation (12) “shows”; other variables (*i.e.*, racial makeup, crime rate, method of census enumeration) controlled for by the regression. This is in some conflict with EK’s theory of the undercount, despite their ingenious argument on p. 56.

The best equation satisfying EK’s current criteria is, in fact, equation (GLS 12). It does not have an intercept. If an intercept is required, the best equation is

$$\begin{aligned} \text{PEP 2-9} &= 1.260 + 2.609 \text{ CC} + .109 \text{ min} + .0262 \text{ conv} - .190 \text{ pov} + \text{residual} \\ &\quad (2.1) \quad (2.9) \quad (5.1) \quad (4.1) \quad (-3.1) \\ \text{rms residual} &= 1.56. \end{aligned}$$

(CC is an indicator for central cities.) Thus, EK cannot have selected their variables quite the way they say they did.

Again, pov comes in with a significant negative coefficient. Within a central city, there are only two variables: min and pov. The equation says that among minority neighborhoods, the poorer they are, the easier they are to count.

(Equation (2) in EKT is a different GLS regression, with covariance matrix $s^2\mathbf{I} + \mathbf{K}$ rather than \mathbf{K} ; s^2 is the estimated value of σ^2 , and \mathbf{K} is the sample-based covariance matrix of the raw undercounts. See equations (1-6) in our paper.)

Our point (p. 15) was that EK could not infer the model from the data; the switch to GLS does not really help them. EK say (p. 57),

“The real issue is how much the actual estimates obtained from the different regression equations vary. The answer to that, as we have shown above, is that the undercounts do not differ substantially.”

That identifies one real issue, out of many. (Another is the impact of variable selection on nominal variances; see Fay 1992). However, if EK are returning to their position of 1986, that they can adjust subareas, then variable selection will matter:

Table C

RMS residuals from regression equations for PEP 2-9 and PEP 10-8.
Explanatory variables include percent minority, percent conventionally enumerated,
and either the crime rate or the percent urban.

	Ordinary Least Squares		Generalized Least Squares	
	Crime Rate	Percent Urban	Crime Rate	Percent Urban
PEP 2-9	1.53	1.54	1.60	1.57
PEP 10-8	1.35	1.33	1.39	1.35

“For the 66 areas in the study, the choice of variables has some impact on the adjustments, but not a major one since both sets of variables span essentially the same column space. On the other hand, when extrapolating to subareas, the choice of variables matters a lot. [F and N, p. 9].

We turn next to Table 7, and recompute it using GLS. As Table C shows, for GLS as well as OLS, percent urban is a better variable than the crime rate; and PEP 10-8 is better than PEP 2-9. EK’s reasons for excluding 10-8 do not survive inspection.

The simulation in Table 9 comes out very much the same way, whether you select the variables by OLS or GLS. Table D repeats the simulation in Table 10, fitting by GLS. EK are right: urb comes in a little less often, CC noticeably more often. Still, urb beats three of EK’s variables (if by a whisker, in the case of MU). Furthermore, negative signs are hardly uncommon in the GLS runs, with paradoxical consequences noted above.

Table D

A simulation experiment on variable selection.
PEP 2-9 is taken as “truth”; percent urban (Urb) is permitted as an explanatory variable.
The table shows the number of times each variable is entered, and the average of its coefficient
(over the times it enters); 100 data sets were generated. In both regimes, coefficients
must be significant; with GLS, negative values are permitted.

Variable	Ordinary Least Squares		Generalized Least Squares	
	No. of Times Entered	Average Coefficient	No. of Times Entered	Average Coefficient
CC	17	2.954	34	2.922
Min	82	0.071	92	0.084
Crime	53	0.053	40	0.055
Conv	93	0.028	94	0.028
Ed	5	0.085	11	−0.099
Pov	1	0.135	25	−0.212
Lang	17	0.315	5	0.417
MU	0	*****	18	−0.048
Urb	23	0.060	19	0.053

Notes: CC is an indicator for central cities; Min, the percentage of minorities; Crime, the crime rate; Conv, the percentage who were conventionally enumerated; Ed, the percentage with no high school degree; Pov, the percentage below the poverty line; Lang, the percentage who have difficulty with English; MU, the percentage living in multiple-unit housing.

Final Comments

EK say (p. 58),

“Freedman and Navidi hold the adjustment data to a higher standard than the unadjusted data. They take on faith, and contrary to decades of Census Bureau evidence, that the unadjusted data are accurate, and they do not seem to be concerned with an evident pattern of bias across areas.”

That is wrong on all counts. Our article begins with a discussion of errors in the census, their variation across areas, and the resource implications. However, we think that the censuses of 1980 and 1990, with overall accuracy estimated in the range 98% to 99%, were considerable achievements. Management skills have been learned from two centuries of experience, and there was dedicated work by hundreds of thousands of ordinary citizens. These censuses were not perfect, but they were very good of their kind.

Ericksen and Kadane have a novel statistical method which, they say, will improve on the census. Our response is this. Show us. Show us not by the standards of physics on the one hand or ESP research on the other, but by the standards of rational argument. Two court cases and countless journal articles later, we find that Ericksen and Kadane cannot make the argument. But readers will judge for themselves.

ADDITIONAL REFERENCES

- BERAN, R., and 12 other statisticians (1988). Statement on census adjustment. U.S. House of Representatives, Subcommittee on Census and Population, hearing of March 3.
- BRYANT, B. (1992). Memoranda to Michael Darby and Mark Plant. Reprinted in *Dividing the Dollars*, a report of the U.S. Senate Committee on Government Affairs, S Prt 102-83, Washington, DC, 78-86.
- CRESSIE, N. (1987). Comment. *Journal of the American Statistical Association*, 82, 980-983.
- FAY, R.E. (1992). Inferences for small domain estimates from the 1990 Post Enumeration Survey. Technical report, Bureau of the Census, Washington, DC.
- FREEDMAN, D.A., KLEIN, S.P., SACKS, J., SMYTH C.A., and EVERETT, C.G. (1991). Ecological regression and voting rights. *Evaluation Review* 15, 673-711.
- KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. Technical Report, Bureau of the Census, Washington, DC.
- MADANSKY, A. (1986). Comment. *Statistical Science*, 1, 28-30.
- MULRY, M. (1991). 1990 Post Enumeration Survey Evaluation Project P16. Total Error in PES Estimates for Evaluation Post Strata. Technical Report, U.S. Bureau of the Census, Washington, DC.
- U.S. DEPARTMENT OF COMMERCE (1991a). Office of the Secretary. Decision on Whether or Not a Statistical Adjustment of the 1990 Decennial Census of Population Should Be Made for Coverage Deficiencies Resulting in an Overcount or Undercount of the Population; Explanation. Report dated July 15, Washington, DC.
- U.S. DEPARTMENT OF COMMERCE (1991b). Press Release CB91-221, dated 6/13/91.
- U.S. DEPARTMENT OF COMMERCE (1991c). Press Release CB91-222, dated 6/13/91.
- WOLTMAN, H.F., *et al.* (1991). Loss function evaluation. Technical report on project P16, U.S. Bureau of the Census, Washington, DC.

REML Estimation in Empirical Bayes Smoothing of Census Undercount

NOEL CRESSIE¹

ABSTRACT

One way to assess the undercount at subnational levels (*e.g.* the state level) is to obtain sample data from a post-enumeration survey, and then smooth those data based on a linear model of explanatory variables. The relative importance of sampling-error variances to corresponding model-error variances determines the amount of smoothing. Maximum likelihood estimation can lead to oversmoothing, so making the assessment of undercount over-reliant on the linear model. Restricted maximum likelihood (REML) estimators do not suffer from this drawback. Empirical Bayes prediction of undercount based on REML will be presented in this article, and will be compared to maximum likelihood and a method of moments by both simulation and example. Large-sample distributional properties of the REML estimators allow accurate mean squared prediction errors of the REML-based smoothers to be computed.

KEY WORDS: Linear model; Maximum likelihood; Restricted maximum likelihood; Variance components.

1. INTRODUCTION

Although a census attempts to carry out a complete enumeration of the population, for various reasons the final tallies are inaccurate. Census personnel, from its director down to the thousands of temporary enumerators, are part of a mammoth task whose accuracy relies on everyone doing their jobs to perfection.

Moreover, events that are beyond human control (*e.g.* weather, natural disaster) must stay within expected limits. Clearly, in a country the size of the U.S.A. (in terms of both population and geography), many opportunities arise to give an imperfect census count. But size is not the only problem; **heterogeneity** of both population and geography gives a **differentially** imperfect count.

The inaccuracies are typically expressed in terms of undercount, so that a negative value implies an overcount. Suppose the U.S.A. is divided into $i = 1, \dots, n$ areas (*e.g.* states, including Washington DC). In the i -th area, let T_i be the true (unknown) count and C_i be the census count. Then the undercount, expressed as a percentage of the true count, is defined as,

$$U_i \equiv \{ (T_i - C_i) / T_i \} 100. \quad (1.1)$$

The problem of **differential** undercount is a serious one when census counts are used to apportion political power and revenue to areas and subareas. (Further discussion of these issues can be found in Ericksen and Kadane 1985, Freedman and Navidi 1986 and Cressie 1988). States like California, Texas, and New York would gain much from adjusting for undercount, *i.e.* from replacing C_i with $F_i C_i$, where F_i is an **adjustment factor**.

The correct adjustment to use is,

$$F_i = T_i / C_i, \quad (1.2)$$

¹ Noel Cressie, Department of Statistics, Iowa State University, Ames, IA, U.S.A. 50011.

which is related to undercount by,

$$F_i = \{1 - U_i/100\}^{-1}.$$

As it stands, (1.2) is not helpful for adjustment, since the true count T_i is unknown. To obtain extra information that will allow F_i to be estimated, the U. S. Census Bureau conducts a post-enumeration survey (PES) that determines whether people in the PES were or were not counted in the census (*e.g.* Wolter 1986). The survey consists of several hundred thousand households, yielding "raw" adjustment factors $\{Y_i : i = 1, \dots, n\}$ that are in need of smoothing.

Assume that, given F_i ,

$$Y_i \sim \text{Gau}(F_i, \delta_i^2), \quad (1.3)$$

i.e. Y_i has, conditional on F_i , a Gaussian distribution with mean F_i and variance δ_i^2 . Adding the further assumption of independence, one obtains,

$$\underline{Y} \sim \text{Gau}(\underline{F}, \Delta), \quad (1.4)$$

where $\underline{Y} \equiv (Y_1, \dots, Y_n)'$, $\underline{F} \equiv (F_1, \dots, F_n)'$, and Δ is the $n \times n$ diagonal matrix $\text{diag}\{\delta_1^2, \dots, \delta_n^2\}$.

Now assume that,

$$\underline{F} \sim \text{Gau}(X\underline{\beta}, \Gamma(\tau^2)), \quad (1.5)$$

where X is an $n \times p$ matrix of explanatory variables, $\underline{\beta}$ is a $p \times 1$ vector of (unknown) coefficients of the linear model, $\Gamma(\tau^2)$ is an $n \times n$ diagonal matrix:

$$\Gamma(\tau^2) \equiv \tau^2 D \quad (1.6)$$

and $D \equiv \text{diag}\{1/C_1, \dots, 1/C_n\}$. The heteroskedastic model (1.5) and (1.6) is discussed at considerable length in Cressie (1990). It is intuitively sensible that the adjustment factor, for an area whose population is large, has a smaller variance; Cressie (1989) provides both a Bayesian and a frequentist justification for this intuition.

Another way to write the model (1.4) and (1.5) is:

$$\underline{Y} = X\underline{\beta} + \underline{v} + \underline{\epsilon}, \quad (1.7)$$

where the $n \times 1$ vectors \underline{v} and $\underline{\epsilon}$ are statistically independent, $\underline{v} \sim \text{Gau}(\underline{0}, \Gamma(\tau^2))$, and $\underline{\epsilon} \sim \text{Gau}(\underline{0}, \Delta)$. Now, assuming that $\delta_1^2, \dots, \delta_n^2$ are calculated using sampling-variance formulas appropriate for the PES sampling frame, the only parameters left to estimate are $\underline{\beta}$ and τ^2 . Thus, the two variance components Δ and $\Gamma(\tau^2)$ only contribute one unknown parameter, namely τ^2 . It is worth noting that the methods developed in this article can be easily generalized beyond this simple variance-components problem. The general linear model is considered in Section 3.

In Section 2, the Bayes predictor and the empirical Bayes predictor of \underline{F} will be given. Estimation of $\underline{\beta}$ is straightforward, but there are several possible ways τ^2 could be estimated. Section 3 presents maximum likelihood (m.l.), method-of-moments, and restricted maximum likelihood (REML) approaches. The effect of estimation of τ^2 , on mean squared prediction errors, is investigated in Section 4. Section 5 compares the approaches by simulation and by example, and Section 6 presents conclusions and a discussion.

2. EMPIRICAL BAYES PREDICTION

In this article, the true population of any small area is considered to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. Therefore, statistical models for undercount are **conditional** on the observed census counts. The model (1.4), (1.5), and (1.6) has been introduced in Section 1, and will be assumed throughout Sections 2, 3, and 4.

Using a matrix analogue of squared-error loss, the optimal predictor is $E(\underline{F} | \underline{Y})$ (Cressie 1990), which is,

$$\underline{p}^*(\underline{Y}) \equiv \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underline{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}X\beta \quad (2.1)$$

and the mean-squared-prediction-error matrix is,

$$E\{(\underline{F} - \underline{p}^*(\underline{Y}))(\underline{F} - \underline{p}^*(\underline{Y}))'\} = \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}\Gamma(\tau^2). \quad (2.2)$$

For the loss matrix, $L(\underline{F}, \underline{p}) \equiv (\underline{F} - \underline{p})(\underline{F} - \underline{p})'$, (2.1) is easily seen to be a **Bayes** predictor of \underline{F} . In reality, β and τ^2 are unknown and so (2.1) is not a statistic (*i.e.* is not a function only of the data). The proper Bayesian approach would be to put further priors and hyperpriors on all unknown parameters. (This solution to the conundrum of unknown parameters is sometimes called hierarchical Bayes, and demands a prior knowledge of process variability that many scientists do not feel they have. Nevertheless, noninformative priors and hyperpriors, particularly, often yield sensible estimators.) Often the posterior distributions are analytically intractable. Should the model and prior be specified according to their conditional distributions, the Gibbs sampler could be used to obtain, numerically, all required marginal and joint distributions (*e.g.* Gelfand and Smith 1990).

An alternative approach, the one taken in this article, is to treat all parameters, except \underline{F} , as fixed but unknown, and to use the data \underline{Y} to estimate them. This approach is called **empirical Bayes**. Although a parametric (conjugate) prior is assumed in this article, one could also work with a nonparametric prior (*e.g.* Laird and Louis 1987).

Suppose now that β is unknown, but that τ^2 in (1.6) is (for the moment) known. Again, using the matrix analogue of squared-error loss, the optimal linear unbiased predictor is obtained by substituting the generalized-least-squares estimator:

$$\hat{\beta} \equiv \{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underline{Y}$$

into (2.1), yielding

$$\begin{aligned} \hat{\underline{p}}(\underline{Y}; \tau^2) &= \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underline{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\} \\ &\quad X\{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underline{Y} \equiv \Lambda(\tau^2)\underline{Y} \end{aligned} \quad (2.3)$$

(Cressie 1990). The mean-squared-prediction-error matrix is,

$$\begin{aligned} M_1(\tau^2) &\equiv E\{(\underline{F} - \hat{\underline{p}}(\underline{Y}; \tau^2))(\underline{F} - \hat{\underline{p}}(\underline{Y}; \tau^2))'\} \\ &= \Lambda(\tau^2)\Delta\Lambda(\tau^2)' + (\Lambda(\tau^2) - I)\Gamma(\tau^2)(\Lambda(\tau^2) - I)'. \end{aligned} \quad (2.4)$$

More realistically, τ^2 is also unknown. An **empirical Bayes predictor** is obtained by substituting an estimator $\hat{\tau}^2$ into $\Lambda(\tau^2)$ to yield,

$$\hat{\underline{p}}(\underline{Y}; \hat{\tau}^2) = \Lambda(\hat{\tau}^2) \underline{Y}. \quad (2.5)$$

It is easy to see that when $\hat{\tau}^2$ is the maximum likelihood estimator of τ^2 , then (2.5) is the maximum likelihood estimator of the Bayes predictor.

The predictor (2.5) was suggested by Ericksen and Kadane (1985) (and criticized by Freedman and Navidi 1986). Incidentally, the form of their predictors may look different to (2.1), (2.3), and (2.5), but they are in fact identical upon using the identity: $A(A + B)^{-1}B = (A^{-1} + B^{-1})^{-1}$, where A and B are square matrices such that A , B , and $A + B$ have inverses.

By substituting $\hat{\tau}^2$ into (2.4), an estimator of the mean-squared-prediction-error matrix:

$$M_1(\hat{\tau}^2) \equiv \Lambda(\hat{\tau}^2)\Delta\Lambda(\hat{\tau}^2)' + (\Lambda(\hat{\tau}^2) - I)\Gamma(\hat{\tau}^2)(\Lambda(\hat{\tau}^2) - I)' \quad (2.6)$$

is obtained. Since (2.6) does not take into account the estimation of τ^2 in $\hat{\underline{p}}(\underline{Y}; \hat{\tau}^2)$, it is likely to be a biased estimator of $E\{(\underline{F} - \hat{\underline{p}}(\underline{Y}; \hat{\tau}^2))(\underline{F} - \hat{\underline{p}}(\underline{Y}; \hat{\tau}^2))'\}$. Further discussion of this important issue is given in Section 4.

Having obtained $\hat{\underline{\beta}}$ and $\hat{\tau}^2$, model diagnostics can be computed to check the fit of the estimated model. For example, a quantile-quantile plot, of the standardized residuals $(\Delta + \Gamma(\hat{\tau}^2))^{-1/2}(\underline{Y} - X\hat{\underline{\beta}})$ against expected order statistics from a unit Gaussian distribution, was used to show no obvious lack of fit of the model used in Section 5. A more complete discussion of model diagnostics is given in Section 6.

3. ESTIMATION OF VARIANCE-MATRIX PARAMETERS

In this section, the general linear model,

$$\underline{Y} \sim \text{Gau}(X\underline{\beta}, \Sigma(\gamma)), \quad (3.1)$$

will be assumed, where γ is a $k \times 1$ vector of variance-matrix parameters. In particular, the model given by (1.4), (1.5), and (1.6) yields,

$$\Sigma(\gamma) = \Delta + \Gamma(\tau^2), \quad (3.2)$$

where γ consists of only one parameter, τ^2 .

For γ known, estimation of $\underline{\beta}$ is straightforward:

$$\hat{\underline{\beta}}(\gamma) \equiv (X' \Sigma(\gamma)^{-1} X)^{-1} X' \Sigma(\gamma)^{-1} \underline{Y}. \quad (3.3)$$

More realistically, γ is unknown and has to be estimated; substitution of that estimator into (3.3) then yields an estimated generalized least squares estimator of $\underline{\beta}$. In the rest of this section, three different methods of estimating γ will be considered.

3.1 Maximum Likelihood Estimation

The negative log likelihood of $\underline{\beta}$ and γ is:

$$\begin{aligned} L(\underline{\beta}, \gamma) = & (n/2)\log(2\pi) + (1/2)\log(|\Sigma(\gamma)|) + \\ & (1/2)(\underline{Y} - X\underline{\beta})' \Sigma(\gamma)^{-1} (\underline{Y} - X\underline{\beta}). \end{aligned} \quad (3.4)$$

Minimization of this function yields maximum likelihood (m.l.) estimates $\hat{\beta}_{m\ell}$ and $\hat{\gamma}_{m\ell}$. The difficult part of this minimization involves finding $\hat{\gamma}_{m\ell}$. The Gauss-Newton (scoring) algorithm is given *inter alia* by Harville (1977) and Mardia and Marshall (1984) and is repeated here for notational completeness.

Define,

$$\begin{aligned}\Sigma_i(\gamma) &\equiv \partial \Sigma(\gamma) / \partial \gamma_i; i = 1, \dots, k, \\ \Sigma^i(\gamma) &\equiv \partial \Sigma^{-1}(\gamma) / \partial \gamma_i = - \Sigma(\gamma)^{-1} \Sigma_i(\gamma) \Sigma(\gamma)^{-1}; i = 1, \dots, k,\end{aligned}\quad (3.5)$$

the $k \times 1$ vector \underline{L}_γ to have i -th element:

$$(\underline{L}_\gamma)_i \equiv (1/2) \text{tr}(\Sigma(\gamma)^{-1} \Sigma_i(\gamma)) + (1/2) (\underline{Y} - X\beta)' \Sigma^i(\gamma) (\underline{Y} - X\beta), \quad (3.6)$$

and the $k \times k$ matrix J_γ to have (i,j) -th element:

$$(J_\gamma)_{ij} \equiv (1/2) \text{tr}(\Sigma(\gamma)^{-1} \Sigma_i(\gamma) \Sigma(\gamma)^{-1} \Sigma_j(\gamma)). \quad (3.7)$$

Then,

$$\gamma^{(\ell+1)} = \gamma^{(\ell)} - (J_\gamma^{(\ell)})^{-1} \underline{L}_\gamma^{(\ell)}, \quad (3.8)$$

where $J_\gamma^{(\ell)}$ and $\underline{L}_\gamma^{(\ell)}$ denotes J_γ and \underline{L}_γ , respectively, evaluated at $\gamma = \gamma^{(\ell)}$ and $\beta = \hat{\beta}(\gamma^{(\ell)})$.

When γ consists of only τ^2 in (1.6), the algorithm (3.8) is particularly straightforward. In the simulations and example given in Section 5, the starting value

$$\begin{aligned}(\tau^2)^{(0)} &\equiv \{1/(n-p)\} (\underline{Y} - X(X'D^{-1}X)^{-1}X'D^{-1}\underline{Y})' D^{-1} \\ &\quad (\underline{Y} - X(X'D^{-1}X)^{-1}X'D^{-1}\underline{Y}),\end{aligned}\quad (3.9)$$

was used. Then (3.8) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - \left\{ (1/2) \sum_{i=1}^n 1/(C_i \delta_i^2 + (\tau^2)^{(\ell)})^2 \right\}^{-1} L_\tau^{(\ell)}; \ell = 0, 1, \dots, \quad (3.10)$$

where

$$\begin{aligned}L_\tau^{(\ell)} &= (1/2) \sum_{i=1}^n 1/(C_i \delta_i^2 + (\tau^2)^{(\ell)}) \\ &\quad - (1/2) \{ \underline{Y} - X\hat{\beta}((\tau^2)^{(\ell)}) \}' \text{diag}\{C_i/(C_i \delta_i^2 + (\tau^2)^{(\ell)})^2\} \{ \underline{Y} - X\hat{\beta}((\tau^2)^{(\ell)}) \}. \quad (3.11)\end{aligned}$$

Iterating (3.8) to convergence yields the m.l. estimator $\hat{\gamma}_{m\ell}$, which upon substitution into (3.3) yields the m.l. estimator $\hat{\beta}(\hat{\gamma}_{m\ell})$. Under appropriate regularity conditions (e.g. Mardia and Marshall 1984) $(\hat{\beta}(\hat{\gamma}_{m\ell})', \hat{\gamma}_{m\ell}')'$ is approximately multivariate Gaussian, with mean $(\beta', \gamma')'$ and asymptotic variance matrix,

$$\begin{bmatrix} (X' \Sigma(\gamma)^{-1} X)^{-1} & 0 \\ 0 & J_\gamma^{-1} \end{bmatrix}; \quad (3.12)$$

when γ consists of only τ^2 in (1.6), the matrix (3.12) becomes,

$$\begin{bmatrix} (X' \Sigma(\tau^2)^{-1} X)^{-1} & 0 \\ 0 & \left\{ (\frac{1}{2}) \sum_{i=1}^n 1/(C_i \delta_i^2 + \tau^2)^2 \right\}^{-1} \end{bmatrix}. \quad (3.13)$$

In practice, estimated variances and covariances are obtained by evaluating (3.12) at the m.l. estimate $\hat{\gamma}_{ml}$.

3.2 Method-of-Moments Estimation

There is no single method-of-moments estimator of γ , but the general idea is to match low-order moments of data with corresponding empirical moments. If only first- and second-order moments are used, it is clear that the Gaussian assumption in (3.1) is not needed.

Let U be a positive-definite symmetric matrix. Consider the weighted regression estimator, $\hat{\beta}_U \equiv (X' U^{-1} X)^{-1} X' U^{-1} Y$, and the weighted residuals,

$$e_U \equiv U^{-1/2} (I - X(X' U^{-1} X)^{-1} X' U^{-1}) Y. \quad (3.14)$$

Then, straightforward matrix algebra shows that,

$$E(e'_U e_U) = \text{tr}(\Sigma(\gamma) \Pi_U), \quad (3.15)$$

where $\Pi_U \equiv U^{-1} - U^{-1} X(X' U^{-1} X)^{-1} X' U^{-1}$. Assuming that $\Sigma(\gamma) = \Delta + \gamma_1 \Gamma_1 + \dots + \gamma_k \Gamma_k$, where Γ_i 's are known, one obtains,

$$\sum_{i=1}^k \gamma_i \text{tr}(\Gamma_i \Pi_U) = E(e'_U e_U) - \text{tr}(\Delta \Pi_U).$$

Choice of k different U_j ; $j = 1, \dots, k$ (e.g. U_1, U_1^2, \dots, U_1^k) yields k equations in k unknowns:

$$\sum_{i=1}^k \gamma_i \text{tr}(\Gamma_i \Pi_{U_j}) = e'_{U_j} e_{U_j} - \text{tr}(\Delta \Pi_{U_j}); j = 1, \dots, k, \quad (3.16)$$

which can be solved for $\hat{\gamma}_1, \dots, \hat{\gamma}_k$. It is important to check that the solution $\hat{\gamma}$ is in the parameter space $\{\gamma: \sum_{i=1}^k \gamma_i \Gamma_i \text{ is positive-definite}\}$.

When γ consists of only τ^2 in (1.6), only one matrix U in (3.16) is needed. Previous under-count predictors have based their estimate of τ^2 on $U = I$ (Ericksen and Kadane 1985;

Freedman and Navidi 1986; Ericksen, Kadane and Tukey 1989), but a small sensitivity study for the heteroskedastic model (1.6) suggested a better estimator.

Choose $U_\alpha = \Delta + \Gamma(\alpha)$ in (3.15) to mimic the model (1.7). Then, when $\alpha = \tau^2$ (the true value), Fay and Herriot (1979) show that

$$E(\underline{e}'_{U_\alpha} \underline{e}_{U_\alpha}) = n - p, \quad (3.17)$$

where n is the number of areas, p is the number of regressors in the matrix X (e.g. $p = 3$ for the selected model in Section 5), and \underline{e}_{U_α} is the standardized residual defined by (3.14). Thus, the proposed method-of-moments estimator of τ^2 is the value of α for which

$$\underline{e}'_{U_\alpha} \underline{e}_{U_\alpha} = n - p, \quad (3.18)$$

which can be solved using a Newton-Raphson iterative method or a simple bisection method; call the resulting estimator $\hat{\tau}_{mm}^2$.

Fay and Herriot (1979) note that the difference between $\hat{\tau}_{mm}^2$ and $\hat{\tau}_{ml}^2$ is manifest in how an area with small δ_i^2 is weighted in the estimation procedure; $\hat{\tau}_{ml}^2$ gives relatively more weight to the squared residuals for such an area than does $\hat{\tau}_{mm}^2$. Based on this weighting property, and a small simulation study of bias, Cressie (1990) expressed a preference for $\hat{\tau}_{mm}^2$ over $\hat{\tau}_{ml}^2$. However, asymptotically, $\hat{\tau}_{ml}^2$ is fully efficient and has an accessible distribution theory. Lack of any (asymptotic) distributional results for $\hat{\tau}_{mm}^2$ causes its own set of problems, such as how to make inference on τ^2 , and how to carry out mean-squared-prediction-error corrections in Section 4. A more satisfactory estimator, with better bias properties than the m.l. estimator, is developed below.

3.3 Restricted Maximum Likelihood Estimation

The problem is to find a suitable estimator of the variance-matrix parameters γ in (3.1). The method of restricted maximum likelihood (REML), developed originally by Patterson and Thompson (1971, 1974), applies maximum likelihood to error contrasts rather than to the data themselves. (Rao (1979) calls this method MML, marginal maximum likelihood, in the context of estimation of variance components. Recently, some authors have also called it residual maximum likelihood, although they have retained the abbreviation REML.) A linear combination $\underline{q}'\underline{Y}$ is called an error contrast if $E(\underline{q}'\underline{Y}) = 0$, for all $\underline{\beta}$ and γ ; thus, $\underline{q}'\underline{Y}$ is an error contrast if and only if $\underline{q}'\underline{X} = \underline{0}'$.

Let $\underline{W} = \underline{A}'\underline{Y}$ represent a vector of $(n - p)$ linearly independent error contrasts; i.e. the $(n - p)$ columns of \underline{A} are linearly independent and $\underline{A}'\underline{X} = \underline{0}$. Under the Gaussian assumption (3.1), $\underline{W} \sim \text{Gau}(\underline{0}, \underline{A}'\underline{\Sigma}(\gamma)\underline{A})$, which does not depend on $\underline{\beta}$. Thus, the negative log likelihood function is,

$$\begin{aligned} L_W(\gamma) = & ((n - p)/2)\log(2\pi) + (\frac{1}{2})\log(|\underline{A}'\underline{\Sigma}(\gamma)\underline{A}|) + \\ & (\frac{1}{2})\underline{W}'(\underline{A}'\underline{\Sigma}(\gamma)\underline{A})^{-1}\underline{W}. \end{aligned}$$

If another set of $(n - p)$ linearly independent contrasts were used to define \underline{W} , the new negative log likelihood function would differ from $L_W(\gamma)$ only by an additive constant (Harville 1974). Indeed, for the \underline{A} that satisfies $\underline{A}\underline{A}' = \underline{I} - \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'$ (and $\underline{A}'\underline{A} = \underline{I}$),

$$\begin{aligned} L_W(\gamma) = & ((n - p)/2)\log(2\pi) - (\frac{1}{2})\log(|\underline{X}'\underline{X}|) + (\frac{1}{2})\log(|\underline{\Sigma}(\gamma)|) + \\ & (\frac{1}{2})\log(|\underline{X}'\underline{\Sigma}(\gamma)^{-1}\underline{X}|) + (\frac{1}{2})\underline{Y}'\underline{\Pi}(\gamma)\underline{Y}, \end{aligned} \quad (3.19)$$

where $\Pi(\gamma) \equiv \Sigma(\gamma)^{-1} - \Sigma(\gamma)^{-1}X(X'\Sigma(\gamma)^{-1}X)^{-1}X'\Sigma(\gamma)^{-1}$; see Harville (1974). A REML estimate of γ , denoted $\hat{\gamma}_{re}$, is obtained by minimizing (3.19) with respect to γ . The distinction between REML and m.l. estimation becomes important when p is large relative to n . The REML method was originally proposed to estimate variance-component parameters: Numerical algorithms (Harville 1977), robust adaptations (Fellner 1986), and distribution theory (Cressie and Lahiri 1991) have been developed in this context. Kitanidis (1983) and Zimmerman (1989) give computational details for producing an iterative minimization of (3.19).

Harville (1974) provides a Bayesian justification for REML by assuming a noninformative prior for β , which is statistically independent of γ , and showing that the marginal posterior density of γ is proportional to (3.19) multiplied by the prior for γ . When that prior is noninformative, REML estimates correspond to marginal MAP (maximum *a posteriori*) estimates. Thus, in the situation where noninformative prior distributions for β and γ are independent, REML can be seen as a compromise between m.l. and Bayes estimation with squared error loss. In the case of model (1.4), (1.5) and (1.6), the latter would yield a Bayes estimate, $\int_0^\infty \tau^2 \exp\{-L_W(\tau^2)\} d\tau^2$, which can be obtained equivalently by **averaging** τ^2 , weighted by the *full* likelihood, $\exp\{-L(\beta, \tau^2)\}$. On the other hand, m.l. yields as an estimate of τ^2 the value $\hat{\tau}_{ml}^2$ obtained by **maximizing** the **full** likelihood. REML averages the full likelihood over β but maximizes the resulting (restricted) likelihood over τ^2 .

Maximum likelihood estimation of τ^2 tends to be biased towards zero because the likelihood, as a function of τ^2 , is skewed to the right. When normalized to integrate to one, the mean of such a function is generally larger than its mode (*e.g.* Groeneveld and Meeden 1977). The m.l. estimate is based on the profile of the likelihood surface of β and τ^2 , and this favors smaller values of τ^2 . (In contrast, REML is obtained by first integrating the likelihood over β and then maximizing the result over τ^2 . Notice that Bayesians might advocate further integration over τ^2 .)

Although the Bayesian interpretation of REML helps to explain its properties, $\hat{\gamma}_{re}$ also has the obvious frequentist interpretation of being an estimator based on restricted information.

Minimization of (3.19) with respect to γ can proceed by any of the gradient algorithms. Recall,

$$\underline{W} = A' \underline{Y} \quad (3.20)$$

and suppose A satisfies:

$$AA' = I - X(X'X)^{-1}X', \text{ and } A'A = I.$$

For the moment, focus all attention on the $(n - p)$ “data” \underline{W} ; their joint distribution depends only on γ , and the associated negative log (restricted) likelihood is $L_W(\gamma)$ given by (3.19).

Define the $k \times 1$ vector \underline{M}_γ to have i -th element:

$$(\underline{M}_\gamma)_i \equiv \partial L_W(\gamma) / \partial \gamma_i = (\frac{1}{2}) \text{tr}\{\Pi(\gamma) \Sigma_i(\gamma)\} - (\frac{1}{2}) \underline{Y}' \Pi(\gamma) \Sigma_i(\gamma) \Pi(\gamma) \underline{Y}, \quad (3.21)$$

and the $k \times k$ matrix G_γ to have (i, j) -th element:

$$(G_\gamma)_{ij} \equiv E(\partial^2 L_W(\gamma) / \partial \gamma_i \partial \gamma_j) = (\frac{1}{2}) \text{tr}\{\Pi(\gamma) \Sigma_i(\gamma) \Pi(\gamma) \Sigma_j(\gamma)\}, \quad (3.22)$$

where $\Pi(\gamma)$ is given below (3.19) and $\Sigma_i(\gamma)$ is defined by (3.5). (The expressions (3.21) and (3.22) were obtained by Harville 1977.) Then, the Gauss-Newton (scoring) algorithm to find $\hat{\gamma}_{re}$ is:

$$\gamma^{(\ell+1)} = \gamma^{(\ell)} - (G_\gamma^{(\ell)})^{-1} M_\gamma^{(\ell)}, \quad (3.23)$$

where $G_\gamma^{(\ell)}$ and $M_\gamma^{(\ell)}$ denote G_γ and M_γ , respectively, evaluated at $\gamma = \gamma^{(\ell)}$.

When γ consists of only τ^2 in (1.6), the algorithm (3.23) is particularly straightforward. In the simulations and example given in Section 5, the starting value (3.9) was used. Then (3.23) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - (G_\tau^{(\ell)})^{-1} M_\tau^{(\ell)}, \quad (3.24)$$

where

$$M_\tau = (\frac{1}{2}) \text{tr} \{ \Pi(\tau^2) D \} - (\frac{1}{2}) Y' \Pi(\tau^2) D \Pi(\tau^2) Y, \quad (3.25)$$

$$G_\tau = (\frac{1}{2}) \text{tr} \{ \Pi(\tau^2) D \Pi(\tau^2) D \}, \quad (3.26)$$

$$\Pi(\tau^2) = \Sigma(\tau^2)^{-1} - \Sigma(\tau^2)^{-1} X (X' \Sigma(\tau^2)^{-1} X)^{-1} X' \Sigma(\tau^2)^{-1}, \quad (3.27)$$

are evaluated at $\tau^2 = (\tau^2)^{(\ell)}$. Also, recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$ and $D = \text{diag}\{1/C_1, \dots, 1/C_n\}$.

Iterating (3.23) to convergence yields the REML estimator $\hat{\gamma}_{r\ell}$. It has been proved by Cressie and Lahiri (1991) that $\hat{\gamma}_{r\ell}$ is approximately multivariate Gaussian, with mean γ and asymptotic variance matrix,

$$G_\gamma^{-1}. \quad (3.28)$$

When γ consists of only τ^2 in (1.6), the matrix (3.28) becomes a scalar,

$$[(\frac{1}{2}) \text{tr} \{ \Pi(\tau^2) D \Pi(\tau^2) D \}]^{-1}. \quad (3.29)$$

In practice, estimated variances and covariances are obtained by evaluating (3.28) at $\gamma = \hat{\gamma}_{r\ell}$. Furthermore, the normalized (estimated) generalized least squares estimator, $\hat{\beta}(\hat{\gamma}_{r\ell})$ should be approximately Gaussian with asymptotic variance matrix, $(X' \Sigma(\gamma) X)^{-1}$.

4. IMPROVED ESTIMATION OF MEAN SQUARED PREDICTION ERRORS

In what is to follow, I shall be concerned with the effect, on prediction, of estimation of γ in $\Sigma(\gamma)$ given by (3.1). Generalizing (1.5) to,

$$F \sim \text{Gau}(X\beta, \Gamma(\gamma)), \quad (4.1)$$

it is clear that

$$\Sigma(\gamma) = \Delta + \Gamma(\gamma). \quad (4.2)$$

In principle, Δ could also depend on unknown parameters (in, *e.g.* a model for sampling variances) and the results of this section are equally applicable. The optimal linear unbiased predictor is,

$$\begin{aligned}\hat{\underline{p}}(\underline{Y}; \underline{\gamma}) &= \Gamma(\underline{\gamma})(\Delta + \Gamma(\underline{\gamma}))^{-1}\underline{Y} + \{I - \Gamma(\underline{\gamma})(\Delta + \Gamma(\underline{\gamma}))^{-1}\} \\ &\quad X\{X'(\Delta + \Gamma(\underline{\gamma}))^{-1}X\}^{-1}X'(\Delta + \Gamma(\underline{\gamma}))^{-1}\underline{Y} \equiv \Lambda(\underline{\gamma})\underline{Y}.\end{aligned}\quad (4.3)$$

Then, the mean-squared-prediction-error matrix of $\hat{\underline{p}}(\underline{Y}; \underline{\gamma})$, denoted $M_1(\underline{\gamma})$, is given by,

$$M_1(\underline{\gamma}) = \Lambda(\underline{\gamma})\Delta\Lambda(\underline{\gamma})' + (\Lambda(\underline{\gamma}) - I)\Gamma(\underline{\gamma})(\Lambda(\underline{\gamma}) - I)'. \quad (4.4)$$

In reality, $\underline{\gamma}$ is unknown and has to be estimated by $\hat{\underline{\gamma}}$, say. The empirical Bayes predictor of \underline{F} is then $\hat{\underline{p}}(\underline{Y}; \hat{\underline{\gamma}})$, given by (4.3) with $\underline{\gamma} = \hat{\underline{\gamma}}$. In this case, $M_1(\underline{\gamma})$ is an inappropriate measure of the predictor's precision; one should use instead,

$$M_2(\underline{\gamma}) = E\{(\underline{F} - \hat{\underline{p}}(\underline{Y}; \hat{\underline{\gamma}}))(\underline{F} - \hat{\underline{p}}(\underline{Y}; \hat{\underline{\gamma}}))'\}. \quad (4.5)$$

It is the risk matrix (4.5), or an estimate of it, that should be given, along with the predictor $\hat{\underline{p}}(\underline{Y}; \hat{\underline{\gamma}})$. However, $M_1(\hat{\underline{\gamma}})$ is typically reported; hence, one should ask what inaccuracies result from using $M_1(\hat{\underline{\gamma}})$ and whether a more appropriate estimator of $M_2(\underline{\gamma})$ is available.

Now, under the assumptions (4.1) and (4.2) (Gaussianity is important here) and provided $\hat{\underline{\gamma}}$ is an even and translation invariant function of the data, the results of Harville (1985) can be used to establish that $M_2(\underline{\gamma}) - M_1(\underline{\gamma})$ is non-negative-definite. (An estimator is even if $\hat{\underline{\gamma}}(\underline{Y}) = \hat{\underline{\gamma}}(-\underline{Y})$ and is translation invariant if $\hat{\underline{\gamma}}(\underline{Y} + X\underline{\lambda}) = \hat{\underline{\gamma}}(\underline{Y})$, for any $p \times 1$ vector $\underline{\lambda}$.) When $\underline{\gamma}$ consists of only τ^2 in (1.6), the estimators $\hat{\tau}_{m\ell}^2$, $\hat{\tau}_{mm}^2$ and $\hat{\tau}_{r\ell}^2$ are all even and translation invariant. Intuitively, estimation of the unknown parameters $\underline{\gamma}$ leads to larger mean squared prediction errors; the result above quantifies this intuition.

But, there is another potential source of bias due to the fact that $M_1(\hat{\underline{\gamma}})$, not $M_1(\underline{\gamma})$, is used to estimate the risk matrix. Suppose that $\hat{\underline{\gamma}}$ is chosen to yield an unbiased estimator of the variance matrix of $(\underline{Y}', \underline{F}')'$, which most would agree is a desirable property. Then the results of Eaton (1985) and Zimmerman and Cressie (1991) can be used to establish that $M_1(\underline{\gamma}) - E(M_1(\hat{\underline{\gamma}}))$ is non-negative-definite. (The proof relies on a multivariate version of Jensen's inequality and on the fact that $\hat{\underline{p}}(\underline{Y}; \underline{\gamma})$, which can be written as $\Lambda(\underline{\gamma})\underline{Y}$, minimizes the risk matrix over all linear unbiased predictors.)

Upon writing,

$$\begin{aligned}M_2(\underline{\gamma}) - M_1(\hat{\underline{\gamma}}) &= \{M_2(\underline{\gamma}) - M_1(\underline{\gamma})\} + \{M_1(\underline{\gamma}) - E(M_1(\hat{\underline{\gamma}}))\} + \\ &\quad \{E(M_1(\hat{\underline{\gamma}})) - M_1(\hat{\underline{\gamma}})\},\end{aligned}\quad (4.6)$$

the results above establish that underestimation of $M_2(\underline{\gamma})$ comes from two sources. Even if an expression for $M_2(\underline{\gamma})$ were known, it is likely that $M_2(\hat{\underline{\gamma}})$ would be biased for $M_2(\underline{\gamma})$, further illustrating the inherent difficulty in estimating mean squared prediction errors.

A remedy has been suggested by Prasad and Rao (1990), based on asymptotic expansions of $M_2(\underline{\gamma})$. Consider prediction of undercount in the i -th area, and let $[M_2(\underline{\gamma})]_{ii}$ and $[M_1(\hat{\underline{\gamma}})]_{ii}$ denote the (i, i) -th elements of the risk matrices $M_2(\underline{\gamma})$ and $M_1(\hat{\underline{\gamma}})$, respectively. Then formal application of Prasad and Rao's proposal yields the estimator of $[M_2(\underline{\gamma})]_{ii}$,

$$[M_2(\underline{\gamma})]_{ii}^* \equiv [M_1(\hat{\underline{\gamma}})]_{ii} + 2\text{tr}\{A_{ii}(\hat{\underline{\gamma}})B(\hat{\underline{\gamma}})\}; i = 1, \dots, n. \quad (4.7)$$

In (4.7), $A_{ii}(\gamma)$ is a $k \times k$ matrix given by,

$$A_{ii}(\gamma) = \text{var}\{\partial \hat{p}_i(Y; \gamma) / \partial \gamma\} \quad (4.8)$$

and $B(\gamma)$ is a matrix that equals or approximates the $k \times k$ matrix,

$$E\{(\hat{\gamma} - \gamma)(\hat{\gamma} - \gamma)'\}. \quad (4.9)$$

For m.l. estimation,

$$B(\gamma) = J_\gamma^{-1}, \quad (4.10)$$

where J_γ is given by (3.7), and for REML estimation,

$$B(\gamma) = G_\gamma^{-1}, \quad (4.11)$$

where G_γ is given by (3.22).

Kass and Steffey (1989) give approximations (to the conditional variance) that are similar in spirit to (4.7), for probability distributions that are not necessarily Gaussian. However, their approach requires independent replications, which is not a feature of the distributions specified by (3.1).

Should small areas be aggregated, it is important to have an approximately unbiased estimator of all elements of $M_2(\gamma)$. It is not difficult to generalize (4.7) to,

$$[M_2(\gamma)]_{ij}^* = [M_1(\hat{\gamma})]_{ij} + 2\text{tr}\{A_{ij}(\hat{\gamma})B(\hat{\gamma})\}; i, j = 1, \dots, n,$$

where $A_{ij}(\gamma) \equiv \text{cov}\{\partial \hat{p}_i(Y; \gamma) / \partial \gamma, \partial \hat{p}_j(Y; \gamma) / \partial \gamma\}$. Prasad and Rao (1990) show that, to the same order of magnitude, $A_{ij}(\gamma)$ can be replaced by $\text{cov}\{\partial p_i^*(Y) / \partial \gamma, \partial p_j^*(Y) / \partial \gamma\}$, where $p^*(Y)$ is given by (2.1); these latter derivatives can be simpler to calculate.

When γ consists of only τ^2 in (1.6), calculation of $B(\gamma)$ is straightforward; see (3.13) and (3.26). Now, consider

$$\text{var}(\partial \hat{p}(Y; \tau^2) / \partial \tau^2) = (\partial \Lambda(\tau^2) / \partial \tau^2) \Sigma(\tau^2) (\partial \Lambda(\tau^2) / \partial \tau^2)', \quad (4.12)$$

where $\Lambda(\tau^2)$ is given by (2.3). In terms of $\Pi(\tau^2)$ defined by (3.27), and Δ defined by (1.4),

$$\Lambda(\tau^2) = I - \Delta \Pi(\tau^2). \quad (4.13)$$

Thus, (4.12) can be calculated from (4.13) using the relationships (3.4) and (3.5). Then, $A_{ii}(\tau^2)$ given by (4.8) is the (i, i) -th element of,

$$\Delta (\partial \Pi(\tau^2) / \partial \tau^2) \Sigma(\tau^2) (\partial \Pi(\tau^2) / \partial \tau^2)' \Delta', \quad (4.14)$$

where

$$\begin{aligned} \partial \Pi(\tau^2) / \partial \tau^2 = & -\Sigma(\tau^2) D \Sigma(\tau^2) \{I - X(X' \Sigma(\tau^2)^{-1} X)^{-1} X' \Sigma(\tau^2)^{-1}\} - \\ & \Sigma(\tau^2)^{-1} X(X' \Sigma(\tau^2)^{-1} X)^{-1} \{X' \Sigma(\tau^2)^{-1} D \Sigma(\tau^2)^{-1} X\} \\ & (X' \Sigma(\tau^2)^{-1} X)^{-1} X' \Sigma(\tau^2)^{-1} + \Sigma(\tau^2)^{-1} X(X' \Sigma(\tau^2)^{-1} X)^{-1} \\ & X' \Sigma(\tau^2)^{-1} D \Sigma(\tau^2)^{-1}; \end{aligned} \quad (4.15)$$

recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$, and $D = \text{diag}\{1/C_1, \dots, 1/C_n\}$.

The estimator of mean squared prediction error, $[M_2(\tau^2)]_{ii}^*$, is conjectured to be approximately unbiased (Prasad and Rao's 1990, results were obtained for a more specific model than is considered here). It is obtained by bringing together the relations (4.7), (4.14) and (4.10) for m.l. estimation, or (4.7), (4.14) and (4.11) for REML estimation. This estimator will be compared to the often-reported estimator $[M_1(\hat{\tau}^2)]_{ii}$, in Section 5, using 1980 U. S. Census and Post Enumeration Survey data.

5. A COMPARISON OF ESTIMATORS BY EXAMPLE AND BY SIMULATION

5.1 Example

The PEP 3-8 data from the 1980 Post Enumeration Survey, for the $n = 51$ states of the USA (including Washington, DC) are used to illustrate the empirical Bayes approach. These data are presented in Cressie (1989, Table 1, "Total" columns) and the variances $\delta_1^2, \dots, \delta_{51}^2$ in (1.3) are obtained from Cressie's "Total" column labeled $MSE^{1/2}$ (whose squared entries will be denoted MSE_1, \dots, MSE_{51}). Using the relation $F_i = \{1 - U_i/100\}^{-1}$ and the δ -method, $\delta_i^2 \approx (Y_i)^4 (MSE_i)/10^4$. Eight explanatory variables, given by Ericksen, Kadane and Tukey (1989), were collapsed to the 51 states (from 66 small areas that included cities, rest of states and states). The explanatory variables are:

1. Minority percentage.
2. Crime rate.
3. Poverty percentage.
4. Percentage with language difficulty.
5. Education.
6. Housing.
7. Proportion of population in any of 16 prespecified central cities.
8. Percentage conventionally counted in the census.

To find a subset of these variables that provides a good model for undercount, I used the selection method of Ericksen, Kadane and Tukey (1989), but weighted the data proportionally to the square roots of the small areas' census counts. The variables selected were 1 (minority) and 5 (education), as well as the constant term. Henceforth, in this paper, these three variables will be the only ones considered in the linear model; *i.e.* only regression coefficients β_0, β_1 and β_5 will be fit.

Under the model (1.4), (1.5) and (1.6), the unknown parameters are $\underline{\beta}$ and τ^2 . From the scoring algorithm (3.8), the m.l. estimate of τ^2 is:

$$\hat{\tau}_{ml}^2 = 47.32,$$

while from the scoring algorithm (3.23), the REML estimate of τ^2 is:

$$\hat{\tau}_{rl}^2 = 58.53.$$

This illustrates a phenomenon observed from the realizations of a simulation presented below, namely, that $\hat{\tau}_{ml}^2 < \hat{\tau}_{rl}^2$; an intuitive explanation is given in Section 3.3. (Parenthetically, Cressie (1990), obtained $\hat{\tau}_{mm}^2 = 94.96$, but no general inequality between it, m.l., and REML is apparent.)

From the formulas in Section 3, the following estimates (with estimated standard errors in parentheses) were obtained:

m.l.	REML
$\hat{\beta}_0 = 1.03227 \ (0.00708)$	$\hat{\beta}_0 = 1.03246 \ (0.00724)$
$\hat{\beta}_1 = 0.0006878 \ (0.0001402)$	$\hat{\beta}_1 = 0.0006941 \ (0.0001436)$
$\hat{\beta}_5 = -0.001070 \ (0.000231)$	$\hat{\beta}_5 = -0.001078 \ (0.000236)$
$\hat{\tau}^2 = 47.32 \ (32.87)$	$\hat{\tau}^2 = 58.53 \ (38.1).$

Notice that there is very little difference between the two sets of estimates, except for that of τ^2 . Upon using the m.l. and REML estimates in $\hat{p}_i(Y; \hat{\tau}^2)$ given by (2.5), $[M_1(\hat{\tau}^2)]_{ii}$ given by (2.6), and $[M_2(\tau^2)]_{ii}^*$ given by (4.7); $i = 1, \dots, n$, small-area predictors and estimated root mean squared prediction errors are obtained. Table 1 shows the results for the $n = 51$ states; also shown in the table are the raw undercount data Y_i , the fitted linear model $(X\hat{\beta})_i$, and the weight,

$$w_i \equiv \hat{\tau}^2 / (C_i \delta_i^2 + \hat{\tau}^2), \tag{5.1}$$

such that

$$\hat{p}_i(Y; \hat{\tau}^2) = w_i Y_i + (1 - w_i) (X\hat{\beta})_i; i = 1, \dots, 51. \tag{5.2}$$

Notice that w_i for REML is consistently larger than w_i for m.l., which is intuitively sensible since $\hat{\tau}_{ml}^2$ has a notoriously large, negative bias. Thus, REML estimation of τ^2 results in less weight on the model term $(X\hat{\beta})_i$, but in a way so that the effect of estimation of τ^2 can be incorporated.

It is interesting to notice that one pays a price for using REML; its root mean squared prediction errors are consistently larger. This is not surprising, since we know that (asymptotically) m.l. is 100% efficient. Further, notice that the improved root mean squared prediction error, $\sqrt{[M_2(\tau^2)]_{ii}^*}$, is between 1% and 9% larger than $\sqrt{[M_1(\hat{\tau}^2)]_{ii}}$.

With regard to prediction, one can assess the importance of m.l. versus REML estimation of τ^2 by computing the weighted sum of squares,

$$\sum_{i=1}^{51} \{ \hat{p}_i(Y; \hat{\tau}_{ml}^2) - \hat{p}_i(Y; \hat{\tau}_{rl}^2) \}^2 C_i = 15.$$

When compared to,

$$\sum_{i=1}^{51} (Y_i - 1)^2 C_i = 70,421$$

and

$$\sum_{i=1}^{51} \{ Y_i - \hat{p}_i(Y; \hat{\tau}_{ml}^2) \}^2 = 26,033,$$

Table 1: Columns, from left to right, show the 51 states according to a three-letter identifier, their raw undercounts $\{ Y_i \}$, their model fits $\{ (X\hat{\beta})_i \}$, their weights $\{ w_i \}$ given by (5.1), their predictors (5.2) (headed F12), their root mean squared prediction errors $\{ \sqrt{[M_1(\hat{\tau}^2)]_{ii}} \}$ (headed RMPE1), and their improved root mean squared prediction errors $\{ \sqrt{[M_2(\tau^2)]_{ii}^*} \}$ (headed RMPE2). Table is given over the page.

Table 1

STATE	Y	REML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1431	1.0026	0.00439	0.00453
aka	1.0288	1.0175	0.4767	1.0229	0.00896	0.00976
arz	1.0204	1.0158	0.0742	1.0162	0.00487	0.00500
ark	0.9895	0.9962	0.1398	0.9953	0.00541	0.00562
cal	1.0307	1.0225	0.0682	1.0231	0.00322	0.00327
col	1.0033	1.0199	0.1926	1.0167	0.00473	0.00495
con	0.9886	1.0079	0.1029	1.0059	0.00435	0.00451
del	0.9938	1.0107	0.4571	1.0030	0.00739	0.00811
fla	1.0144	1.0120	0.0785	1.0122	0.00289	0.00295
gga	0.9955	1.0046	0.1639	1.0031	0.00391	0.00403
hai	1.0111	1.0105	0.2785	1.0107	0.00678	0.00730
idh	1.0125	1.0070	0.5627	1.0101	0.00531	0.00579
ill	1.0211	1.0103	0.1170	1.0116	0.00257	0.00265
ind	0.9936	1.0026	0.1413	1.0013	0.00334	0.00349
iow	0.9932	1.0033	0.1478	1.0018	0.00452	0.00475
kan	1.0056	1.0092	0.2215	1.0084	0.00466	0.00496
ky	0.9845	0.9872	0.1519	0.9868	0.00507	0.00524
lou	1.0234	1.0086	0.0263	1.0090	0.00476	0.00480
mne	1.0201	0.9992	0.3703	1.0069	0.00593	0.00645
mld	1.0242	1.0140	0.0712	1.0147	0.00406	0.00415
mas	0.9882	1.0068	0.1945	1.0032	0.00323	0.00341
mch	1.0079	1.0081	0.1601	1.0081	0.00259	0.00271
min	1.0111	1.0049	0.2793	1.0066	0.00359	0.00383
mis	1.0097	1.0086	0.1279	1.0087	0.00557	0.00575
mou	1.0080	1.0010	0.1681	1.0022	0.00350	0.00367
mon	1.0144	1.0059	0.3785	1.0091	0.00699	0.00761
neb	1.0008	1.0071	0.5117	1.0039	0.00441	0.00480
nev	1.0265	1.0151	0.2852	1.0183	0.00744	0.00802
nwh	0.9842	1.0033	0.3080	0.9974	0.00684	0.00740
nwj	1.0130	1.0105	0.0895	1.0107	0.00305	0.00314
nwm	1.0236	1.0256	0.3276	1.0249	0.00611	0.00648
nwy	1.0166	1.0119	0.0807	1.0123	0.00243	0.00247
noc	1.0118	0.9998	0.0748	1.0007	0.00421	0.00430
nod	1.0005	0.9969	0.8931	1.0001	0.00313	0.00324
oho	1.0108	1.0044	0.1273	1.0052	0.00253	0.00263
okl	0.9977	1.0018	0.1625	1.0011	0.00429	0.00451
ore	1.0027	1.0089	0.2833	1.0071	0.00434	0.00464
pen	0.9972	1.0013	0.1475	1.0007	0.00253	0.00263
rhi	1.0089	0.9939	0.4167	1.0001	0.00625	0.00678
soc	1.0632	1.0040	0.0216	1.0053	0.00555	0.00559
sod	1.0008	0.9985	0.7538	1.0002	0.00464	0.00496
ten	0.9717	0.9966	0.0755	0.9947	0.00439	0.00449
tex	1.0037	1.0149	0.0482	1.0144	0.00341	0.00345
uth	1.0040	1.0142	0.4010	1.0101	0.00524	0.00563
vmt	0.9889	1.0018	0.8232	0.9912	0.00454	0.00479
vir	1.0009	1.0058	0.1753	1.0049	0.00338	0.00354
was	1.0142	1.0121	0.1305	1.0123	0.00418	0.00434
wev	0.9942	0.9877	0.1452	0.9887	0.00603	0.00628
wis	1.0173	1.0032	0.2877	1.0073	0.00325	0.00348
wyo	1.0361	1.0127	0.3992	1.0221	0.00882	0.00963
dcl	1.0375	1.0474	0.2191	1.0452	0.01081	0.01125

Table 1 (concluded)

STATE	Y	ML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1190	1.0028	0.00415	0.00427
aka	1.0288	1.0175	0.4241	1.0223	0.00850	0.00933
arz	1.0204	1.0157	0.0608	1.0160	0.00448	0.00459
ark	0.9895	0.9963	0.1161	0.9955	0.00506	0.00525
cal	1.0307	1.0224	0.0559	1.0228	0.00314	0.00319
col	1.0033	1.0198	0.1617	1.0171	0.00446	0.00466
con	0.9886	1.0079	0.0849	1.0063	0.00398	0.00412
del	0.9938	1.0107	0.4050	1.0039	0.00697	0.00771
fla	1.0144	1.0120	0.0644	1.0121	0.00271	0.00276
gga	0.9955	1.0046	0.1368	1.0034	0.00375	0.00385
hai	1.0111	1.0105	0.2378	1.0106	0.00629	0.00679
idh	1.0125	1.0070	0.5099	1.0098	0.00507	0.00559
ill	1.0211	1.0103	0.0967	1.0113	0.00242	0.00248
ind	0.9936	1.0026	0.1174	1.0015	0.00309	0.00323
iow	0.9932	1.0034	0.1230	1.0021	0.00418	0.00438
kan	1.0056	1.0091	0.1870	1.0085	0.00432	0.00460
ky	0.9845	0.9874	0.1264	0.9870	0.00486	0.00502
lou	1.0234	1.0086	0.0214	1.0089	0.00446	0.00449
mne	1.0201	0.9993	0.3222	1.0060	0.00557	0.00608
mld	1.0242	1.0139	0.0583	1.0145	0.00376	0.00384
mas	0.9882	1.0068	0.1634	1.0037	0.00302	0.00319
mch	1.0079	1.0081	0.1335	1.0081	0.00242	0.00252
min	1.0111	1.0049	0.2386	1.0064	0.00339	0.00362
mis	1.0097	1.0085	0.1060	1.0087	0.00526	0.00541
mou	1.0080	1.0011	0.1404	1.0021	0.00326	0.00341
mon	1.0144	1.0059	0.3299	1.0087	0.00656	0.00717
neb	1.0008	1.0071	0.4587	1.0042	0.00420	0.00461
nev	1.0265	1.0150	0.2439	1.0178	0.00692	0.00746
nwh	0.9842	1.0033	0.2646	0.9983	0.00637	0.00691
nwj	1.0130	1.0105	0.0736	1.0106	0.00283	0.00290
nwm	1.0236	1.0254	0.2826	1.0249	0.00582	0.00617
nwy	1.0166	1.0119	0.0663	1.0122	0.00231	0.00235
noc	1.0118	0.9998	0.0614	1.0005	0.00401	0.00408
nod	1.0005	0.9970	0.8710	1.0000	0.00310	0.00324
oho	1.0108	1.0045	0.1055	1.0051	0.00236	0.00245
okl	0.9977	1.0018	0.1356	1.0013	0.00396	0.00416
ore	1.0027	1.0088	0.2421	1.0074	0.00408	0.00436
pen	0.9972	1.0014	0.1227	1.0008	0.00239	0.00248
rhi	1.0089	0.9940	0.3660	0.9995	0.00591	0.00645
soc	1.0632	1.0041	0.0176	1.0051	0.00519	0.00523
sod	1.0008	0.9985	0.7122	1.0002	0.00452	0.00490
ten	0.9717	0.9967	0.0619	0.9951	0.00413	0.00422
tex	1.0037	1.0148	0.0393	1.0144	0.00329	0.00332
uth	1.0040	1.0141	0.3512	1.0105	0.00498	0.00536
vmt	0.9889	1.0019	0.7901	0.9916	0.00445	0.00477
vir	1.0009	1.0058	0.1467	1.0051	0.00317	0.00330
was	1.0142	1.0120	0.1082	1.0123	0.00391	0.00406
wev	0.9942	0.9879	0.1207	0.9886	0.00567	0.00590
wis	1.0173	1.0033	0.2461	1.0067	0.00306	0.00328
wyo	1.0361	1.0127	0.3494	1.0209	0.00829	0.00909
dcl	1.0375	1.0470	0.1849	1.0452	0.01036	0.01078

it is clear that, from a national perspective, prediction is not very sensitive to estimation methods for τ^2 . (Cressie (1990) reaches the same conclusion based on a similar comparison of $\hat{\tau}_{ml}^2$ and $\hat{\tau}_{mm}^2$.) However, from Table 1, it is equally clear that estimated root mean squared prediction errors are considerably more sensitive.

Cressie (1990) gives expressions for the risks of adjusting using $\hat{p}(Y;\tau^2)$ and of not adjusting. When $\hat{\tau}_{rl}^2$ and $\hat{g}(\hat{\tau}_{rl}^2)$ are substituted into those expressions, the risk of adjusting is 3,253, while the risk, of not adjusting is 34,134. That is, not adjusting leads to a 949% increase in risk (provided the model defined by (1.4), (1.5) and (1.6) holds).

5.2 Simulation

To check the asymptotic distribution theory of the REML (and m.l.) estimator of τ^2 , a simulation was carried out on the linear model described in Section 5.1, with parameter values:

$$\beta_0 = 1.0330, \quad \beta_1 = 0.000712, \quad \beta_5 = -0.000110, \quad \tau^2 = 95.00.$$

(5.3)

The simulation,

$$Y \sim \text{Gau}(X\beta, \Delta + \tau^2 D),$$

(5.4)

where Δ is given by (1.4) the same values of $\delta_1^2, \dots, \delta_{51}^2$, as used in Section 5.1 and Cressie in 1990, are used here and D is given by (1.6), was performed 500 times, and each time the estimates, $\hat{\tau}_{ml}^2$, $\hat{\tau}_{mm}^2$, and $\hat{\tau}_{rl}^2$ were computed. (Whenever a negative value was obtained, the estimate was set equal to zero.) The stem-and-leaf plots of the three sets of estimates are presented in Figures 1a, 1b and 1c, respectively. Notice the relatively larger number of zeros for the m.l. estimates (Figure 1a).

Figure 1. Stem-and-leaf plots of estimated variance parameter τ^2 , based on 500 simulations of (5.4): (a) maximum likelihood (Section 3.1), (b) method-of-moments (Section 3.2) and (c) restricted maximum likelihood (Section 3.3).

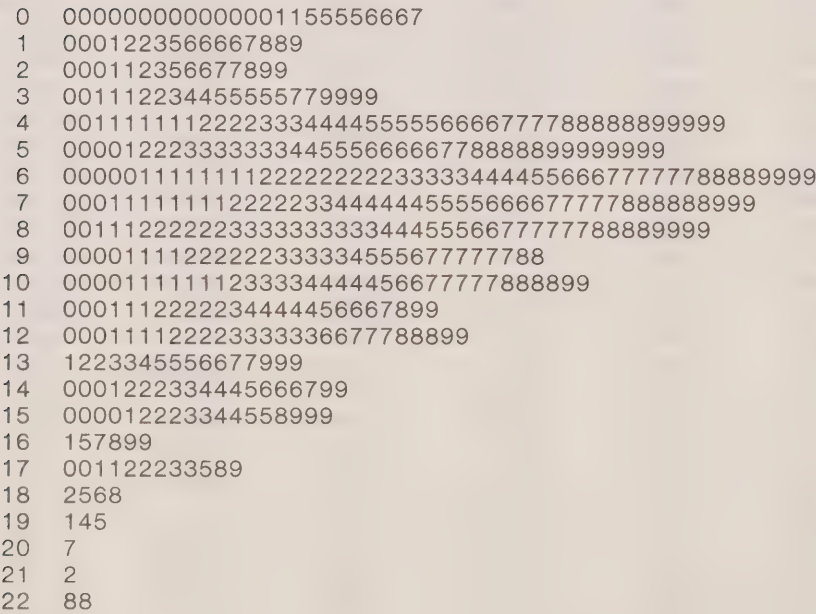


Figure 1a

0 000000377778
1 011113344446679999
2 11144455557778888
3 0000222222233333355555666666888899999
4 12222222446666777777999999
5 00022233333345557777778888
6 000000011133333444444444666667777999999999
7 1111222222444445555555777778888888
8 00000222333333355555556666666688999
9 111222222244444466666777779999
10 0000002222333333355577777888888
11 00000000111113333444444666677789999
12 111112222222444444445777778
13 00002233666888888999
14 11122244667777999
15 00223355888
16 000001133444777799
17 122222245555788
18 00003335566899
19 26799
20 02258
21 37
22 11558
23 5
24 79
25
26
27 5
28 5
29 2
30 78
31 3
32
33 2

Figure 1b

The means (\bar{X}) and standard deviations (S) of the distributions shown in Figure 1 are:

$\hat{\tau}_{m\ell}^2$	$\hat{\tau}_{mm}^2$	$\hat{\tau}_{r\ell}^2$
$\bar{X} = 83.56$	$\bar{X} = 96.85$	$\bar{X} = 94.27.$
$S = 45.65$	$S = 57.46$	$S = 49.17.$

The means should be compared to the true value of $\tau^2 = 95.00$. The bias in $\hat{\tau}_{m\ell}^2$ is apparent; $\hat{\tau}_{r\ell}^2$ has very little bias and has a small advantage over $\hat{\tau}_{mm}^2$. With regard to standard deviations, the advantage of $\hat{\tau}_{r\ell}^2$ over $\hat{\tau}_{mm}^2$ is considerable, but it is at some disadvantage over $\hat{\tau}_{m\ell}^2$. For reasons explained in Section 3.3, that are not all statistical, bias is more of a concern than variance, and so REML estimation of τ^2 should be considered a serious alternative to m.l.

Asymptotic distribution theory for m.l. and REML can be checked from the simulations. (The method of moments is at a disadvantage in that no asymptotic distribution theory is readily available.) Substituting $\tau^2 = 95.00$ into (3.13) yields,

```
0 00000001234777799
1 0012334567789
2 012225556888899
3 112234444556889
4 0013344445555666777888888899
5 000011222333333444445566777788888999
6 00011122222223344444444566667777888899999
7 00000011111122222223333444455567777888899999
8 00000001111222333455555566666777778889
9 0001112222222333333444455555566689999999
10 00000000011122333444555566777888899
11 0001122233344455666677788888899
12 00011111122233344455567789
13 000133334555555556788
14 0001112344445667789
15 00111222344566788
16 0011122223355557999
17 011235556
18 00112566777779
19 117
20 013478
21 123
22 7
23 6
24
25 02
```

Figure 1c

$$\{\text{var}(\hat{\tau}_{m\ell}^2)\}^{1/2} \approx 48.73,$$

which should be compared to $S = 45.65$. Finally, substituting $\tau^2 = 95.00$ into (3.29) yields,

$$\{\text{var}(\hat{\tau}_{r\ell}^2)\}^{1/2} \approx 50.14,$$

which should be compared to $S = 49.17$.

The opportunity also exists to use the simulation to look at “actual” errors of prediction and to assess the performance of $M_1(\hat{\tau}^2)$ and $M_2(\tau^2)^*$. If the parameter values (5.3) were estimated from the original data, then this amounts to a parametric bootstrap.

6. CONCLUSIONS AND DISCUSSION

Model-based prediction of undercount relies on careful checking of model fit. Diagnostic plots based on standardized residuals have already been suggested at the end of Section 2. The standardized BLUP residuals $\{Y_i - \hat{p}_i(Y; \hat{\tau}^2)\} / \{[M(\hat{\tau}^2)]_{ii}\}^{1/2}; i = 1, \dots, n$, also have a role to play. They could either be used in a quantile-quantile plot (e.g. Cressie 1991, p. 225) or, as suggested by Calvin and Sedransk (1991), plotted against $\hat{p}_i(Y; \hat{\tau}^2); i = 1, \dots, n$.

One could also extend the model (1.4) to include an unknown variance-component parameter σ^2 :

$$Y \sim \text{Gau}(F, \sigma^2\Delta), \tag{6.1}$$

where $\Delta = \text{diag}\{\delta_1^2, \dots, \delta_n^2\}$. Upon fitting the more general model (6.1), (1.5) and (1.6), one could then test whether the REML estimate $\sigma_{r\ell}^2$ is significantly different from $\sigma^2 = 1$, which would provide a check on model misspecification. (In this case, REML estimation is recommended over m.l. estimation, since any bias will seriously affect inference on σ^2 .)

Restricted maximum likelihood (REML) estimation of variance-matrix parameters is less likely to lead to empirical Bayes predictors that put too much weight on the regression model (1.5). The price paid is slightly larger mean squared prediction errors. Using asymptotic distribution theory for REML (which is checked by simulation), improved estimators of the mean squared prediction errors can also be obtained. Based on the model (1.4), (1.5) and (1.6), it can be concluded that there are accurate and precise ways to make inference on adjustment factors $\{F_i : i = 1, \dots, n\}$; the predictors $\{\hat{p}_i(Y; \hat{\tau}_{r\ell}^2) : i = 1, \dots, n\}$ yield true-count and undercount predictors,

$$T_i^{\text{prd}} = \hat{p}_i(Y; \hat{\tau}_{r\ell}^2) C_i \quad \text{and} \quad U_i^{\text{prd}} = 100\{1 - (\hat{p}_i(Y; \hat{\tau}_{r\ell}^2))^{-1}\}; \quad i = 1, \dots, n,$$

respectively. Their biases and mean-squared prediction errors can be obtained using the δ -method (cf. Cressie 1991, Section 3.2.2).

ACKNOWLEDGEMENTS

The research assistance of Robert Parker is gratefully acknowledged. This research was supported by Joint Statistical Agreement JSA 90-41, between the U. S. Bureau of the Census and Iowa State University. The findings and opinions expressed in this article are of the author alone, and do not necessarily reflect those of the Census Bureau.

REFERENCES

- CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-48.
- CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. In *Proceedings of Bureau of the Census Fourth Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 127-150.
- CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- CRESSIE, N. (1990). Weighted smoothing of estimated undercount. In *Proceedings of Bureau of the Census 1990 Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 301-325.
- CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.
- CRESSIE, N., and LAHIRI, S.N. (1991). The asymptotic distribution of REML estimators. *Statistical Laboratory Preprint 91-20*, Iowa State University, Ames, Iowa.
- EATON, M.L. (1985). The Gauss-Markov Theorem in multivariate analysis. In *Multivariate Analysis - VI*, (Ed. P.R. Krishnaiah). Amsterdam: Elsevier, 177-201.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.
- ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.
- FAY, R.E., III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

- FELLNER, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-17.
- GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- GROENEVELD, R.A., and MEEDEN, G.D. (1977). The mode, median and mean inequality. *American Statistician*, 31, 120-121.
- HARVILLE, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.
- HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.
- HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.
- KASS, R.E., and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84, 717-726.
- KITANIDIS, P.K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19, 909-921.
- LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.
- MARDIA, K.V., and MARSHALL, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.
- MILLER, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics*, 5, 746-762.
- PATTERSON, H.D., and THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.
- PATTERSON, H.D., and THOMPSON, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*. Washington, DC: Biometric Society, 197-207.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, C.R. (1979). MINQE theory and its relation to ML and MML estimation of variance components. *Sankhyā B*, 41, 138-153.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- ZIMMERMAN, D.L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21, 655-672.
- ZIMMERMAN, D.L., and CRESSIE, N. (1991). Mean-squared prediction error in the spatial linear model. *Annals of the Institute of Statistical Mathematics*, 43, forthcoming.

Hierarchical and Empirical Bayes Methods for Adjustment of Census Undercount: The 1988 Missouri Dress Rehearsal Data

**G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI,
L.K. SCHULTZ and J.H. TSAY¹**

ABSTRACT

The present article discusses a model-based approach towards adjustment of the 1988 Census Dress Rehearsal Data collected from test sites in Missouri. The primary objective is to develop procedures that can be used to model data from the 1990 Census Post Enumeration Survey in April, 1991 and smooth survey-based estimates of the adjustment factors. We have proposed in this paper hierarchical Bayes (HB) and empirical Bayes (EB) procedures which meet this objective. The resulting estimators seem to improve consistently on the estimators of the adjustment factors based on dual system estimation (DSE) as well as the smoothed regression estimators.

KEY WORDS: Post Enumeration Survey; Adjustment factors; Dual system estimation; Hierarchical Bayes; Empirical Bayes; Variance components; EBLUP's; Regression estimates; Standard errors.

1. INTRODUCTION

The present article discusses a model-based approach towards adjustment of the 1988 Census dress rehearsal data collected from test sites in Missouri. The main objective behind this exercise is to develop procedures that can be used to model data from the 1990 Census Post Enumeration Survey (PES) in April, 1991, and smooth survey-based estimates of the so-called "raw adjustment factors". These raw adjustment factors which are ratios of estimates of the unknown total population to the corresponding 1990 Census count, are computed at various levels of aggregation (geographic areas such as cities, suburbs, *etc.*) crossed by various demographic categories (such as age, sex, race, *etc.*). The cross-classified categories are called poststrata.

Before proceeding further, a brief historical anecdote is in order. Adjustment of 1980 decennial census counts in the United States has been a topic of heated debate for nearly a decade. Despite the intensive efforts and the massive expenditure incurred by the U.S. Bureau of the Census to achieve near-complete coverage in the 1980 Census, there have been many lawsuits against the Bureau by individual states and cities demanding revision of the reported counts. In one such instance of litigation, by now well-publicized to the Statistics community in the articles of Ericksen and Kadane (1985) and Freedman and Navidi (1986), New York City among others sued the Census Bureau, and many reputed statisticians appeared as expert witnesses on either side. In particular Ericksen and Kadane appeared on the plaintiff's side, and proposed a model-based approach towards the adjustment of census counts. They advocated shrinking the adjustment factors calculated on the basis of the PES data towards some suitable regression model. This approach documented in Ericksen and Kadane (1985) is similar to the one considered in Fay and Herriot (1979) or Morris (1983). Despite criticism of the Ericksen-Kadane approach by some statisticians (most severely by Freedman and

¹ G.S. Datta, University of Georgia, Athens, GA 30602; M. Ghosh, University of Florida, Gainesville, FL 32611; E.T. Huang, C.T. Isaki, L.K. Schultz and J.H. Tsay, U.S. Bureau of the Census, Washington, DC 20233.

Navidi (1986)), most people recognize the importance of the model-based approach for adjustment. Indeed, in this article, barring a few differences in the assumptions, to be pointed out later in section 2, we use the Fay-Herriot or the Ericksen-Kadane model for the analysis of the 1988 Missouri Dress Rehearsal data. A different model-based approach which does not include co-variates is given in Cressie (1989).

A good description of the PES conducted as part of the 1988 Missouri Dress Rehearsal can be found in Childers and Hogan (1990). Hogan and Wolter (1988) discuss the categories of error that occur in a PES and a means of their evaluation. Basically, the PES design consists of a single stage stratified sample of blocks and dual system estimation of the number of persons by poststrata.

In the present article, we begin at the point where a set of estimated raw adjustment factors and their covariances from the PES are available for modelling based on the 1988 Census Dress Rehearsal Data from the Missouri test sites. It is also assumed that a set of possible explanatory variables defined at the poststrata level and to be used in regression are also available. There are two geographic areas under consideration: the city of St. Louis which is a large central city, and East Central Missouri, which is a collection of areas of moderate population size. In defining the poststrata in St. Louis, persons were classified into the following demographic categories: (i) race: white non-hispanic and others, (ii) owners and non-owners (renters) of dwellings, (iii) sex: male and female, (iv) age groups: 0-9, 10-19, 20-29, 30-44, 45-64 and 65 + . This led to a total of $2 \times 2 \times 2 \times 6 = 48$ adjustment factors for St. Louis. In East Central Missouri, the sex and the age-group categories remained the same as in St. Louis, but instead of (i) and (ii), a new category (i)' classifying persons as (a) White non-Hispanic in Tape Address Register (TAR) areas, (b) White non-Hispanic in non-TAR areas, and (c) others in all areas were introduced. For East Central Missouri, a total of $3 \times 2 \times 6 = 36$ adjustment factors were calculated. Thus, a total of 84 adjustment factors were used for modelling. Within each area, estimated adjustment factors were correlated due to the use of a block cluster sampling scheme. This led to a block-diagonal sample covariance matrix of the adjustment factors of dimensions 48×48 and 36×36 corresponding to St. Louis and East Central Missouri, respectively.

In Section 2 of this article, we describe a general model-based method for obtaining smoothed adjustment factors, and the associated standard errors. Both the hierarchical and empirical Bayes methods are used. The EB method can also be regarded as a variance components method (see for example Harville (1985)). The formulas for posterior standard errors associated with the HB estimators are also provided. We may point out here that an EB method when employed naively can lead to serious underestimates of the associated standard errors. This is due to the fact that a naive EB method does not take into account the uncertainty due to estimation of the unknown variance components. However, Kackar and Harville (1984), and Prasad and Rao (1990) have suggested interesting approximations to the estimated mean squared errors (MSE's) of the EB estimators. Following their principle, we have derived formulas for the estimated MSE's in the present context. We have also pointed out in this section how some (though not all) of the criticisms levelled against the Ericksen-Kadane (1985) procedure by Freedman and Navidi (1986) can be avoided in the present context.

In Section 3, we have analyzed the actual data. The sample estimates, the HB estimates, the EB estimates and the regression estimates of the adjustment factors are all provided. Also, the associated standard errors are given. Both the HB method and the EB methods which take into account the uncertainty due to unknown prior parameters stand on par in their performance, and enjoy a clear-cut superiority over the raw estimates as well as the regression estimates in reducing the estimated standard errors.

Finally, some of the technical details of this paper are given in the Appendix.

2. HB AND EB ESTIMATION

This section describes the general HB and EB estimation procedures for certain hierarchical models. The specific application to estimation of adjustment factors is considered in Section 3.

The following hierarchical model is considered:

- I. $Y | \Theta, \beta, \sigma^2 \sim N(\Theta, V)$, where V is a known $m \times m$ positive definite matrix;
- II. $\Theta | \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$;
- III. β and σ^2 are marginally independent with β uniform (R^p) and σ^2 uniform $(0, \infty)$.

The HB analysis is based on I-III. In the absence of precise prior information on β and σ^2 , we prefer the use of diffuse priors in III. We also analyzed the data with the prior *pdf* of σ^2 proportional to σ^{-2} on $(0, \infty)$. The results were quite similar and are not reported. The following theorem is proved.

Theorem 1. Consider the model given in (I) – (III). Write $\Sigma = V + \sigma^2 I$. Suppose $m \geq p + 3$. Then (i) the conditional pdf of Θ given σ^2 and $Y = y$ is $N(GV^{-1}y, G)$, where

$$G = V - V\Sigma^{-1}V + V\Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}V; \quad (2.1)$$

(ii) the conditional pdf of σ^2 given $Y = y$ is

$$f(\sigma^2 | y) \propto |\Sigma|^{-1/2} |X^T\Sigma^{-1}X|^{-1/2} \exp(-1/2 y^T F y), \quad (2.2)$$

where

$$F = \Sigma^{-1} - \Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}. \quad (2.3)$$

The proof of the theorem is deferred to the appendix. Using formulas for conditional expectations and variances, one then gets

$$E(\Theta | y) = E[E(\Theta | \sigma^2, y) | y] = (E(GV^{-1} | y)) y; \quad (2.4)$$

$$V(\Theta | y) = V[E(\Theta | \sigma^2, y) | y] + E[V(\Theta | \sigma^2, y) | y] = V(GV^{-1} y | y) + E(G | y). \quad (2.5)$$

Using (2.2) and (2.3), one obtains $E(\Theta | y)$ and $V(\Theta | y)$ from (2.4) and (2.5) via numerical integration.

The calculations involved in (2.1) – (2.3) can be somewhat simplified when one uses the spectral decomposition theorem for V . Thus, $V = PDP^T$, where $D = \text{Diag}(d_1, \dots, d_m)$, d_i being the eigenvalues of V , and $P = (\xi_1 \dots, \xi_m)$, ξ_i being the corresponding orthonormal eigenvectors. Using the orthogonality of P , one now gets

$$|\Sigma| = |\sigma^2 I + D| = \prod_{i=1}^m (\sigma^2 + d_i);$$

$$\Sigma^{-1} = P(\sigma^2 I + D)^{-1}P^T;$$

$$X^T \Sigma^{-1} X = (P^T X)^T (\sigma^2 I + D)^{-1} (P^T X);$$

$$F = P(\sigma^2 I + D)^{-1}P^T - P(\sigma^2 I + D)^{-1}(P^T X) \times$$

$$[(P^T X)^T (\sigma^2 I + D)^{-1} (P^T X)]^{-1} (P^T X) (\sigma^2 I + D)^{-1}.$$

The actual numerical integration over σ^2 which needs evaluation of the integrand at different values of σ^2 , is somewhat simplified since P and X are known and $\sigma^2 I + D$ is a diagonal matrix.

Next we consider EB estimation. Then, one does not use III. First a Bayes estimator, *i.e.* the posterior mean of Θ is obtained from I and II assuming β and σ^2 to be known. This estimator is given by

$$\begin{aligned} \hat{\Theta}^B &= E(\Theta | Y, \beta, \sigma^2) \\ &= (V^{-1} + \sigma^{-2}I)^{-1}(V^{-1}Y + \sigma^{-2}X\beta) \\ &= \Sigma^{-1}(\sigma^2 Y + VX\beta). \end{aligned} \quad (2.6)$$

The corresponding posterior variance is given by

$$V(\Theta | Y, \beta, \sigma^2) = (V^{-1} + \sigma^{-2}I)^{-1} = V - V\Sigma^{-1}V.$$

However, in practice, β and σ^2 are unknown, and are estimated via the maximum likelihood method from the marginal distribution of Y which is $N(X\beta, \Sigma)$. These MLE's are denoted by $\hat{\beta}$ and $\hat{\sigma}^2$, where $\hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$, $\hat{\Sigma} = V + \hat{\sigma}^2 I$. Substituting such estimators of Σ , σ^2 and β in (2.6), an EB estimator of Θ is found as

$$\hat{\Theta}^{EB} = \hat{\Sigma}^{-1}(\hat{\sigma}^2 Y + VX\hat{\beta}) = X\hat{\beta} + \hat{\sigma}^2 \hat{\Sigma}^{-1}(Y - X\hat{\beta}). \quad (2.7)$$

The estimator given in (2.7) is also obtainable as an estimated best linear unbiased predictor (EBLUP). First assume that σ^2 is known, and find the BLUP $\hat{\Theta}^{BLUP} = X\tilde{\beta} + \sigma^2 \Sigma^{-1}(Y - X\tilde{\beta})$ of Θ where $\tilde{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$. Next estimate σ^2 by $\hat{\sigma}^2$, its MLE and correspondingly Σ by $\hat{\Sigma}$. Substitution of $\hat{\sigma}^2$, and $\hat{\Sigma}$ in place of σ^2 and Σ in $\hat{\Theta}^{BLUP}$ results in the EBLUP $\hat{\Theta}^{EB}$.

A naive EB estimator of the variance matrix of $\hat{\Theta}^{EB}$ is $V - V\hat{\Sigma}^{-1}V$. This is a gross underestimation of the variance matrix since uncertainty due to estimation of β and σ^2 is not taken into account. If σ^2 is assumed known, and β is assigned a uniform prior on R^p ($m \geq p + 3$), then the HB estimator of Θ is the same as $\hat{\Theta}^{BLUP}$, and the posterior variance matrix is then $M = V - V\Sigma^{-1}V + V\Sigma^{-1}X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} V$. This implies immediately that $E[(\hat{\Theta}^{BLUP} - \Theta)(\hat{\Theta}^{BLUP} - \Theta)^T] = M$, where expectation is taken over the

joint distribution of Y and Θ given in I and II. Thus, in the Bayesian language, $V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V$ can be interpreted as the excess in the posterior variability due to the uncertainty involved in β , while using the classical terminology, the same phenomenon can be interpreted as the excess in the MSE due to the same uncertainty.

We have the additional problem of tackling unknown σ^2 . The Bayesian method enables us to find the posterior distribution of σ^2 given $Y = y$, while even without introducing a prior for Θ , it is still possible to find an approximation to the MSE of $\hat{\Theta}^{EB}$ by adapting an argument of Kackar and Harville (1984) or Prasad and Rao (1990).

The necessary theorem whose proof is deferred to the Appendix is given below.

Theorem 2. An approximate estimate of MSE of $\hat{\Theta}^{EB}$ is given by

$$\widehat{MSE}(\Theta^{EB}) \doteq V - VKV + (VK^3V) [2(\text{tr}\hat{\Sigma}^{-2})^{-1}], \quad (2.8)$$

where

$$K = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}. \quad (2.9)$$

The third term in the right hand side of (2.8) can be interpreted as the excess in the mean squared error due to uncertainty in estimating σ^2 . A general decomposition of the prediction error is given in Harville (1985).

Although the posterior variances $V(\Theta | y)$ associated with the HB estimator $\hat{\Theta}^{HB}$ of Θ and the estimated MSE of the EB estimator $\hat{\Theta}^{EB}$ of Θ are motivated from two distinct inferential philosophies, one common thread tying the two is that they both attempt to incorporate the uncertainty due to estimation of the model variance. For a better understanding of this, note that writing $K = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$,

$$E[V(\Theta | \sigma^2, y)] = G = V - VKV \quad (2.10)$$

and $E(G | y)$ is approximated by $V - VKV$ which is one of the two terms given in (2.8). Also, $E(\Theta | \sigma^2, y) = GV^{-1}y$, and it can be shown after some simplification that $GV^{-1} = I - VK$. Thus, $V(GV^{-1}y | y) = VV(K | y)V$, and $V(K | y)$ is apparently approximated by $K^3[2(\text{tr}\hat{\Sigma}^{-2})^{-1}]$. However, as evidenced later in the numerical calculations of Section 3, MSE approximation of $\hat{\Theta}^{EB}$ need not match $V(\Theta | y)$ perfectly.

In Ericksen and Kadane (1985) one assumption involved was that of known σ^2 . Freedman and Navidi (1986) insisted on estimation of σ^2 , and we have in Theorems 1 and 2 accounted for this source of uncertainty both in a Bayesian and frequentist way. It should be noted that unlike previous work that addressed the estimation of net undercount of total population at the city and balance of state level, our interests lay in the estimation of adjustment factors at finer levels of detail. Operationally, adjustment at the finer levels allows for considerable savings in terms of time and computer costs as census files need to be used only once. Adjustment models using higher levels of geography would require several passes through the census data because they would require a method of distributing the undercount to lower levels of geography. Finally, correlation in the error structure allows the possibility of a non-diagonal V , another important generalization of the Fay-Herriot (1979) or Ericksen-Kadane (1985) model. Thus, the Freedman-Navidi criticism of lack of correlation across estimated adjustment factors does not hold against the present set up. The remaining main criticism of assuming the components of V to be known, whereas in reality these are sample based estimates, is yet to be resolved. Efforts are now being made to model the components of V as a function of

variables such as the number of sample persons, the initial regression predictor, *etc.* It is hoped that such models will stabilize the estimated variances by reducing their variance.

Along with the HB and EB estimators of Θ , there are also the regression estimators given by $\hat{\Theta}^{\text{REG}} = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$. The associated variance-covariance matrix is given by $M_o - \hat{\sigma}^2(M_o\hat{\Sigma}^{-1} + \hat{\Sigma}^{-1}M_o^T - I)$, where $M_o = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T$.

3. DATA ANALYSIS

Let $Y_i = \text{DSE}_i / \text{Census}_i = \text{adjustment factor } i, i = 1, \dots, 84$, and $Y = (Y_1, \dots, Y_{84})^T$. The set of explanatory variables X is quite large when all possible interactions are considered. To simplify the analysis, experts at the Census Bureau were consulted and a reduced set of 22 potential explanatory variables were considered for modelling purposes. (See Huang *et al.* 1991). The number of potential explanatory variables was also limited by the capability of the computer. The present model was selected using a best subset regression procedure with minimum Mallows' C_p as the criterion over a set of 22 possible explanatory variables. Because the computer software required the input data to be in the ordinary least squares situation, we transformed the dependent and explanatory variables in the usual manner. Also, because σ^2 is unknown, an iterative procedure was used.

As an aside, in selecting explanatory variables in the modelling process of adjustment factors for the 1990 Census, a slightly different procedure was used. In 1990, several explanatory variables were forced into the model and a best subset procedure was used to select additional explanatory variables. The change in procedure was made to counteract the potential for understating σ^2 . (See Isaki *et al.* 1991).

The X matrix obtained via best subsets regression is of the form $X = (1_{84}, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$. All of the explanatory variables in X are obtained from the 1988 Dress Rehearsal Census and defined at the poststrata level, the unit of analysis. 1_{84} is a unit vector; X_2 is the indicator variable for St. Louis; X_3 is the indicator variable for renters or is the proportion of renters for the East Central Missouri poststrata; X_4 through X_7 are indicator variables for age groups 0-9, 10-19, 20-29 and 30-44, respectively; X_8 is an indicator or proportion variable for males aged 20-64 that rent; X_9 is an indicator variable for other males aged 20-64; and X_{10} is an indicator variable for other persons in St. Louis.

Using the above design matrix, we obtained $\hat{\beta} = (.9812, -.0271, .0485, .0699, .0695, .0533, .0386, .0628, .0475, .0778)^T$ and $\hat{\sigma}^2 = .000574$. The EB's or the EBLUP's and the associated approximate standard errors can now be computed using formulas derived in Section 2. For consistency, the HB analysis was also performed with the same X matrix (we do not require $\hat{\beta}$ or $\hat{\sigma}^2$ for that analysis).

In Figures 1 and 2 we plot the estimated adjustment factors and standard errors by poststrata. The first 12 poststrata refer to white non-Hispanic non-owners in St. Louis; poststrata 13-24 refer to all other non-owners in St. Louis; poststrata 25-36 refer to white non-Hispanic owners in St. Louis and poststrata 37-48 refer to all other owners in St. Louis. Poststrata 49-60 refer to white non-Hispanic persons in Tape Address Register (TAR) areas in East Central Missouri; poststrata 61-72 refer to white non-Hispanic persons in non-TAR areas in East Central Missouri; poststrata 73-84 refer to all other persons in East Central Missouri.

Within each group of 12 poststrata, the first six refer to males by age 0-9, 10-19, 20-29, 30-44, 45-64 and 65+. We note in Figure 1 that the raw adjustment factors for the other group tend to be higher than those for the white non-Hispanic except for TAR area in East Central Missouri. The same observation nearly holds in Figure 2 concerning the raw standard errors. In Figure 3 a plot of the estimated standard errors versus the adjustment factors is provided.

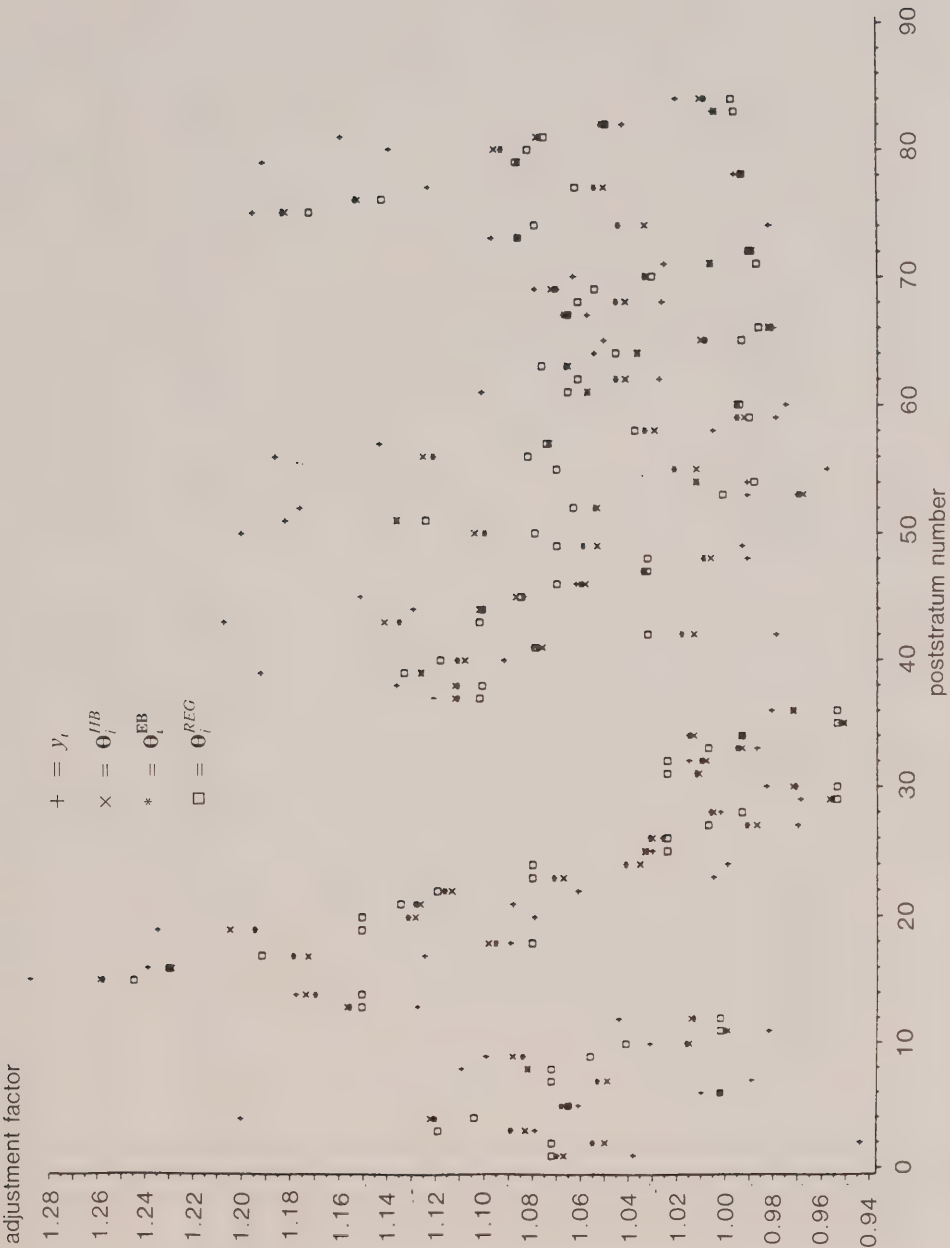


Figure 1. Adjustment Factors by Poststrata.

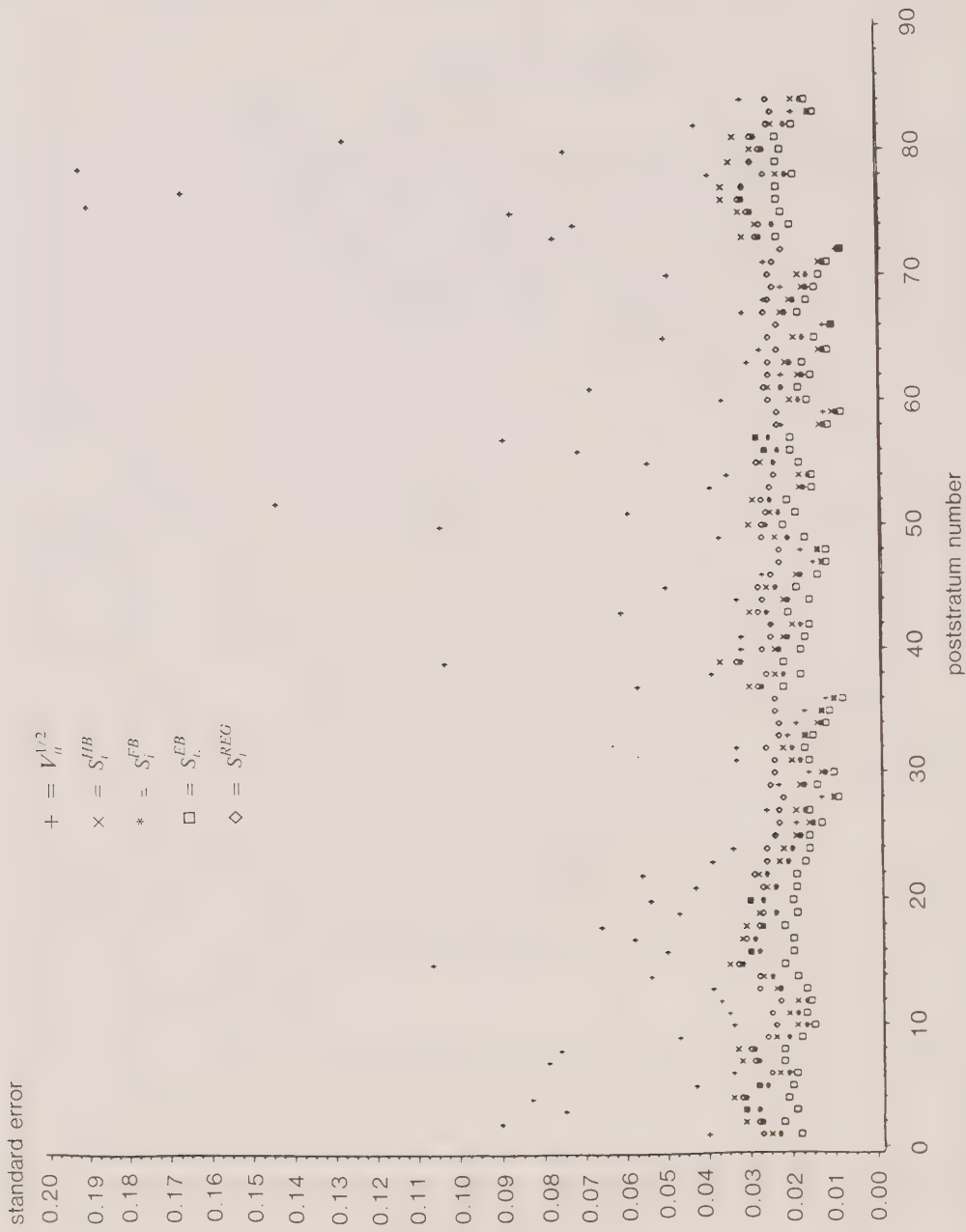


Figure 2. SE of Adjustment Factors by Poststrata.

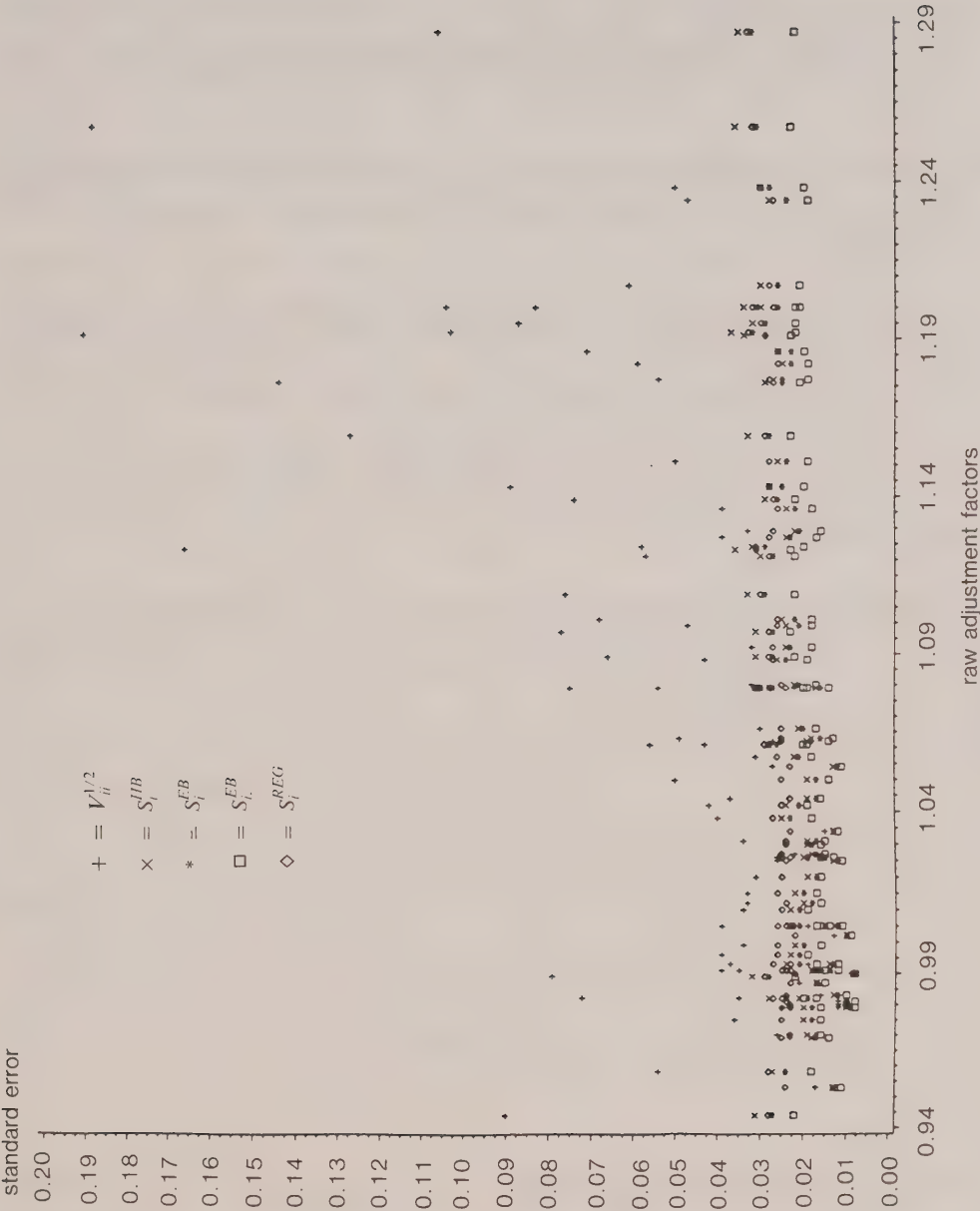


Figure 3. SE of Adjustment Factors by Raw Adjustment Factors.

Figures 1 to 3 lead to several interesting conclusions.

- (1) For every stratum, the estimated standard errors of the HB and the EB estimators of the adjustment factors are much smaller than the standard errors of the raw adjustment factors when compared to the unadjusted DSE's.
- (2) The EB estimators improve on the regression estimators for all the 84 strata by providing reduced estimated standard errors. Although the HB estimators do not improve on the regression estimators for all the strata, the improvement is substantial for most of the strata.
- (3) The data plots demonstrate that the difference between the point estimates $\hat{\theta}_i^{\text{EB}}$ s and $\hat{\theta}_i^{\text{HB}}$ s is quite small. Indeed, the percentage difference is always less than (and most often far less than) 1%.
- (4) The posterior standard errors associated with the HB estimates (s_i^{HB}) are always bigger than the approximate MSE's of the EB estimates (s_i^{EB}). As discussed earlier, the two need not be the same. It is our feeling that the approximate standard errors of the EB estimates are often slight underestimates. However, a comparison of s_i^{EB} and $s_{i.}^{\text{EB}}$ reveals that a naive EB procedure (with associated estimated standard errors $s_{i.}^{\text{EB}}$) can grossly underestimate the estimated standard errors by failing to incorporate uncertainty due to estimation of σ^2 . This deficiency is largely rectified by s_i^{EB} which is based on second order approximations.

At the time of revision of this article, adjustment of the 1990 Decennial Census was completed. The EB estimation procedure was used. Basically, most of the same steps followed in modelling the adjustment factors in the 1988 Dress Rehearsal Census were used. However, there were several differences. In 1990 adjustment, the estimated adjustment factors were modelled by each of four census regions and a special set for Indian reservations. The number of adjustment factors ranged from 12 for the Indian set to 456 in one of the regions. In addition, estimated variances of the raw adjustment factors were smoothed via regression models. Smoothing of the estimated variances tended to reduce large estimated variances and increase small estimated variances. The net effect was an increase in the contribution of the associated adjustment factors with large estimated variances to the EB estimates and vice versa. Other differences were that outlier detection procedures were used in both the variance and adjustment factor smoothing. Finally, the EB estimates at the poststratum level were ratio adjusted to regional total population estimates derived from the raw adjustment factors. The ratio adjusted smoothed factors were then applied to related census population counts at the census block level. The results were then integer rounded by collection of blocks in such a manner that each cell within a block is rounded up or down to an integer and that control totals are off by at most one person.

The procedures used to adjust the 1990 Census counts were pre-specified and the entire operation was conducted under a very tight time schedule. The Bureau of the Census recommended that the 1990 Census adjusted counts be used. A special panel selected by the Secretary of Commerce was evenly divided in this issue. Upon weighing the evidence, the Secretary decided against using the adjusted counts. The issue is now subject to litigation. A current issue is the possible use of adjusted counts for use in postcensal estimation. Research in obtaining better adjusted counts for use in postcensal estimation is currently underway.

APPENDIX - PROOFS OF THE THEOREMS

Proof of Theorem 1. We provide only an outline of the proof. The details appear in Datta *et al.* (1991). The joint (improper) pdf of Y, Θ, β and σ^2 is given by:

$$f(y, \Theta, \beta, \sigma^2) \propto \exp[-1/2(y - \Theta)^T V^{-1}(y - \Theta)] \sigma^{-m} \exp[-1/(2\sigma^2) \|\Theta - X\beta\|^2], \quad (\text{A.1})$$

where $\|\cdot\|$ denotes the Euclidean norm. Writing $P_X = X(X^T X)^{-1}X^T$, $\|\Theta - X\beta\|^2 = [\beta - (X^T X)^{-1}X^T \Theta]^T (X^T X) [\beta - (X^T X)^{-1}X^T \Theta] + \Theta^T (I - P_X) \Theta$.

Now, integrating with respect to β in (A.1), it follows that the joint improper pdf of Y, Θ and σ^2 is

$$f(y, \Theta, \sigma^2) \propto \sigma^{-(m-p)} \exp[-1/2(y - \Theta)^T V^{-1}(y - \Theta) - 1/(2\sigma^2) \Theta^T (I - P_X) \Theta]. \quad (\text{A.2})$$

Next writing $E = V^{-1} + \sigma^{-2}(I - P_X)$, it follows after some simplifications that

$$(y - \Theta)^T V^{-1}(y - \Theta) + \sigma^{-2} \Theta^T (I - P_X) \Theta = (\Theta - E^{-1}V^{-1}y)^T E (\Theta - E^{-1}V^{-1}y) + y^T (V^{-1} - V^{-1}E^{-1}V^{-1})y. \quad (\text{A.3})$$

Hence, the posterior distribution of Θ given σ^2 and $Y = y$ is $N(E^{-1}V^{-1}y, E^{-1})$. Using the familiar matrix inversion formula $(A + BDB^T)^{-1} = A^{-1} - A^{-1}B(D^{-1} + B^T A^{-1}B)^{-1}B^T A^{-1}$ (see for example Exercise 2.9, p. 33 of Rao (1973)), one gets $E^{-1} = G$. This completes the proof of the first part of the Theorem. Next, using (A.3) and integrating with respect to Θ in (A.2), one gets the joint (improper) pdf of Y and σ^2 is

$$f(y, \sigma^2) \propto \sigma^{-(m-p)} |E|^{-1/2} \exp[-(1/2)y^T (V^{-1} - V^{-1}E^{-1}V^{-1})y]. \quad (\text{A.4})$$

Using Exercise 2.4, p. 32 of Rao (1973), it follows that

$$|E| = |V^{-1} + \sigma^2(I - P_X)| = |\Delta| \div |\sigma^2 X^T X|$$

which on simplification reduces to

$$|V^{-1}| |I + \sigma^{-2}V| |X^T (I + \sigma^{-2}V)^{-1}X| \div |X^T X| \propto |I + \sigma^{-2}V| |X^T \Sigma^{-1}X|. \quad (\text{A.5})$$

Also, after some calculations, it follows that

$$V^{-1} - V^{-1}E^{-1}V^{-1} = F. \quad (\text{A.6})$$

The proof of part (ii) of Theorem 1 follows now from (A.4) - (A.6) and noting that $f(\sigma^2 | y) \propto f(\sigma^2, y)$. Note, however that the posterior pdf of σ^2 given $Y = y$ is proper.

Proof of Theorem 2. Once again, only a sketch of the proof is given. The details are available in Datta *et al.* (1991).

Recall

$$\tilde{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y.$$

Define,

$$\hat{\Theta} = X\tilde{\beta} + \sigma^2 \Sigma^{-1} (Y - X\tilde{\beta}).$$

Now, observe that (i) $\hat{\Theta}$ is the best unbiased predictor of Θ (due to normality) for every fixed σ^2 , and (ii) $E(\hat{\Theta}^{EB} - \hat{\Theta}) = \mathbf{0}$ since $\hat{\sigma}^2$ is the MLE of σ^2 (*cf* Kackar and Harville (1984)). Now using Lemma 3.3.1 of Datta (1990), $\hat{\Theta}^{EB} - \hat{\Theta}$ is uncorrelated with $\hat{\Theta}$. Hence,

$$\begin{aligned} E[(\hat{\Theta}^{EB} - \hat{\Theta})(\hat{\Theta}^{EB} - \hat{\Theta})^T] &= \\ E[(\hat{\Theta}^{EB} - \hat{\Theta})(\hat{\Theta}^{EB} - \hat{\Theta})^T] + E[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)^T]. \end{aligned} \quad (A.7)$$

Next, write $\hat{\Theta}^B = X\beta + \sigma^2 \Sigma^{-1} (Y - X\beta)$. Then standard arguments give

$$E[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)^T] = E[(\hat{\Theta} - \hat{\Theta}^B)(\hat{\Theta} - \hat{\Theta}^B)^T] + E[(\hat{\Theta}^B - \Theta)(\hat{\Theta}^B - \Theta)^T]. \quad (A.8)$$

Our previous calculations yield

$$E[(\hat{\Theta}^B - \Theta)(\hat{\Theta}^B - \Theta)^T] = V - V\Sigma^{-1}V. \quad (A.9)$$

Further,

$$E[(\hat{\Theta} - \hat{\Theta}^B)(\hat{\Theta} - \hat{\Theta}^B)^T] = V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V. \quad (A.10)$$

Finally, write $\hat{\Theta} = g(\sigma^2)$ and $\hat{\Theta}^{EB} = g(\hat{\sigma}^2)$. Using first order Taylor approximation, one gets

$$E[(\hat{\Theta}^{EB} - \hat{\Theta})(\hat{\Theta}^{EB} - \hat{\Theta})^T] \doteq E\left[(\hat{\sigma}^2 - \sigma^2)^2 \frac{dg(\sigma^2)}{d\sigma^2} \frac{dg(\sigma^2)^T}{d\sigma^2}\right]. \quad (A.11)$$

Since $g(\sigma^2) = Y - V\Sigma^{-1}[Y - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y]$, using matrix differentiation, techniques, one gets

$$\frac{dg}{d\sigma^2} = V[\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma^{-1}(Y - X\hat{\beta}). \quad (A.12)$$

$$E\left[\frac{dg}{d\sigma^2} \frac{dg^T}{d\sigma^2}\right] = VK\Sigma^{-1}E[(Y - X\hat{\beta})(Y - X\hat{\beta})^T]\Sigma^{-1}KV. \quad (A.13)$$

But, simple algebra gives

$$E[(Y - X\hat{\beta})(Y - X\hat{\beta})^T] = \Sigma - X(X^T\Sigma^{-1}X)^{-1}X^T = \Sigma K\Sigma. \quad (A.14)$$

Hence, from (A.13),

$$E\left[\frac{dg}{d\sigma^2} \frac{dg^T}{d\sigma^2}\right] = VK^3V. \quad (A.15)$$

Using, one more approximation, it follows from (A.11) and (A.15) that

$$E[(\hat{\Theta}^{\text{EB}} - \hat{\Theta})(\hat{\Theta}^{\text{EB}} - \hat{\Theta})^T] \doteq E(\hat{\sigma}^2 - \sigma^2)^2 V K^3 V. \quad (\text{A.16})$$

To estimate $E(\hat{\sigma}^2 - \sigma^2)^2 = \text{MSE}(\hat{\sigma}^2)$, we proceed as follows.

Since $Y \sim N(X\beta, \Sigma)$, write the likelihood function as

$$L(\sigma^2) \propto |\Sigma|^{-1/2} \exp[-1/2(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)]. \quad (\text{A.17})$$

Hence,

$$\frac{d \log L}{d \sigma^2} = -1/2 \frac{d}{d \sigma^2} \log |\Sigma| - 1/2 \frac{d}{d \sigma^2} [(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)]; \quad (\text{A.18})$$

$$\frac{d^2 \log L}{d(\sigma^2)^2} = -1/2 \frac{d^2}{d(\sigma^2)^2} \log |\Sigma| - 1/2 \frac{d^2}{d(\sigma^2)^2} [(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)]. \quad (\text{A.19})$$

As before, denote by d_1, \dots, d_m the eigenvalues of V .

Then, $\log |\Sigma| = \sum_{i=1}^m \log(\sigma^2 + d_i)$. Hence

$$\frac{d^2}{d(\sigma^2)^2} \log |\Sigma| = -\sum_{i=1}^m (\sigma^2 + d_i)^{-2} = -\text{tr}(\Sigma^{-2}). \quad (\text{A.20})$$

Using (A.20) and matrix differentiation, it follows from (A.19) that

$$\frac{d^2 \log L}{d(\sigma^2)^2} = 1/2 \text{tr}(\Sigma^{-2}) - (Y - X\beta)^T \Sigma^{-3}(Y - X\beta). \quad (\text{A.21})$$

Thus,

$$E\left[-\frac{d^2 \log L}{d(\sigma^2)^2}\right] = -1/2 \text{tr}(\Sigma^{-2}) + \text{tr}(\Sigma^{-2}) = 1/2 \text{tr}(\Sigma^{-2}).$$

Approximating $E[(\hat{\sigma}^2 - \sigma^2)^2]$ by

$$\left(E\left[-\frac{d^2 \log L}{d(\sigma^2)^2}\right]\right)^{-1},$$

justifiable by the asymptotic theory of maximum likelihood, one gets, from (A.16),

$$E[(\hat{\Theta}^{\text{EB}} - \hat{\Theta})(\hat{\Theta}^{\text{EB}} - \hat{\Theta})^T] \doteq 2(\text{tr}(\Sigma^{-2}))^{-1} V K^3 V. \quad (\text{A.22})$$

Combining (A.7) - (A.10) and (A.22), one gets

$$\text{MSE}(\hat{\Theta}^{\text{EB}}) \doteq V - V \Sigma^{-1} V + V \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} V + V K^3 V [2(\text{tr} \Sigma^{-2})^{-1}] \quad (\text{A.23})$$

Substitution of $\hat{\Sigma}$ for Σ yields the approximation given in (2.8). This completes the proof of Theorem 2.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation under grant SES 87-13643, "On-Site Research to Improve the Government-Generated Social Science Data Base." The research was conducted at the U.S. Bureau of the Census while the first two authors were participants in the American Statistical Association/Census Bureau Research Program, which is supported by the Census Bureau and through the NSF grant. Any opinions, findings and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of the Census.

REFERENCES

- CHILDERS, D.R., and HOGAN, H. (1990). Results of the 1988 dress rehearsal post enumeration survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 547-552.
- DATTA, G.S. (1990). Bayesian prediction in mixed linear models with applications in small area estimation. Unpublished Ph.D. dissertation. University of Florida.
- DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., and SCHULTZ, L.K. (1990). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress-Rehearsal data. Technical Report No. 376. Department of Statistics, University of Florida.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for Small Places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 3-39.
- HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. *Survey Methodology*, 14, 99-116.
- HUANG, E.T., ISAKI, C.T., and TSAY, J.H. (1991). Modelling PES adjustment factors using 1988 dress rehearsal data. *Proceedings of the Social Statistics Section, American Statistical Association*, to appear.
- ISAKI, C.T., HUANG, E.T., and TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 post enumeration survey. *Proceedings of the Social Statistics Section, American Statistical Association*, to appear.
- KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.
- MORRIS, C. (1983). Parametric Empirical Bayes inference and applications. *Journal of the American Statistical Association*, 78, 47-65.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd. Edition). New York: Wiley.

A Comparison of Some Estimators of a Set of Population Totals

DON ROYCE¹

ABSTRACT

The Population Estimates Program of Statistics Canada has traditionally been benchmarked to the most recent census, with no allowance for census coverage error. Because of a significant increase in the level of undercoverage in the 1986 Census, however, Statistics Canada is considering the possibility of adjusting the base population of the estimates program for net census undercoverage. This paper develops and compares four estimators of such a base population: the unadjusted census counts, the adjusted census counts, a preliminary test estimator, and a composite estimator. A generalization of previously-proposed risk functions, known as the Weighted Mean Square Error (WMSE), is used as the basis of comparison. The WMSE applies not only to population totals, but to functions of population totals such as population shares and growth rates between censuses. The use of the WMSE to develop and evaluate small-area estimators in the context of census adjustment is also described.

KEY WORDS: Census adjustment; Undercoverage; Small area estimation.

1. INTRODUCTION

The Population Estimates Program of Statistics Canada provides a wide variety of detailed information about the characteristics and distribution of the Canadian population during the five-year period between each census. Intercensal estimates of population have many important uses, including the calculation of billions of dollars of transfer payments from the federal to provincial governments, the estimation of important demographic statistics such as birth and mortality rates, the planning of future levels of immigration, and the weighting of current population surveys such as the monthly Labour Force Survey.

Traditionally, the estimates program is based on the most recent census, with no allowance for coverage error. In the 1986 Census, however, undercoverage increased significantly compared to previous censuses, and continued to be distributed unevenly across geographic and demographic groups. This caused considerable disruption to the estimates program and to the many other programs which use population estimates. As a result, a project was initiated in early 1989 to investigate whether, and if so how, the population estimates in the post-1991 Census period should be adjusted for estimated census coverage error. The research described in this paper was conducted as part of this project. For a more general description of the project, see Royce (1992).

It should be noted that only the population estimates would be affected by this adjustment. The 1991 Census data will be published with no adjustment for undercoverage, other than the small adjustments that have traditionally been made to correct for underenumeration of temporary residents and for persons missed because their dwelling was misclassified as vacant. From the technical point of view, however, the question is quite similar to the issue of census adjustment that has been of interest to many statistical agencies in recent years.

Two key questions in the adjustment issue are the degree to which census counts are improved by adjustment, and which adjustment methods are best. In this paper, we compare the accuracy of several different estimators of a set of population totals, using a weighted mean square error as our criterion.

¹ Don Royce, Chief, Census Data Quality Section, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Section 2 introduces the topic by considering the simple case of a single population total. We derive and compare the Mean Square Errors of four possible estimators: the unadjusted census count, the adjusted census count, a preliminary test estimator, and a composite estimator. Section 3 extends the results to multiple population totals and to functions of population totals, such as population shares and growth rates. In Section 4, we consider methods for small-area estimation, specifically, the use of synthetic estimation and a special case of synthetic estimation known as across-the-board adjustment. Section 5 concludes with a description of areas for further research.

In developing the estimators described in this paper, two assumptions were made. First, adjustment must result in estimates that are consistent across all geographic and demographic levels, as well as across time. Users consider it to be essential that parts add up to totals, and that there be no major breaks in the time series of estimates. Second, adjustment will be based on the combined results of Statistics Canada's two coverage measurement studies: the Reverse Record Check, which measures gross undercoverage, and the Overcoverage Study, which measures gross overcoverage. Both studies are subject to sampling errors and non-sampling errors.

2. SINGLE POPULATION TOTAL

We first describe and compare four estimators for the case of a single population total. In comparing the estimators, we use the Mean Square Error (MSE) as our criterion.

- Let: Y be the known census count;
 T be the unknown true population total to be estimated;
 U be the true net undercoverage, *i.e.*, $U = T - Y$;
 \hat{U} be an estimate of U from the coverage measurement studies;
 σ^2 be the variance of \hat{U} ; and
 R be the relative bias of \hat{U} , *i.e.* $R = E(\hat{U})/U - 1$.

In the case of all four estimators, our estimate of T can be written as the census count plus some estimate of U . Thus, the MSE of our estimate of T will be the same as that of the corresponding estimate of U . The MSEs (and the WMSEs in later sections) are taken over hypothetical repetitions of the coverage measurement studies, treating the Census counts as fixed quantities.

2.1 Unadjusted Census Estimator

The unadjusted census estimate of U is zero. It has a bias equal to $-U$ and zero variance. Therefore $MSE(\hat{U}^c) = U^2$.

2.2 Adjusted Census Estimator

The adjusted census estimator of U is \hat{U} . It has a bias of UR and a variance equal to σ^2 . Thus $MSE(\hat{U}^A) = \sigma^2 + U^2R^2$.

2.3 Preliminary Test Estimator

A comparison of the MSEs of the previous two estimators suggests that we would use the adjusted census count in preference to the unadjusted census count whenever

$$\sigma^2 < U^2(1 - R^2). \quad (1)$$

Although the parameters in this inequality are unknown, they can (with the exception of R) be estimated from the coverage measurement studies. This suggests the possibility of using these estimates to develop a statistical test of the hypothesis that the inequality holds. The result of the test is then used to choose which estimator to use (thus the term preliminary test, or pretest, estimator).

Specifically, assume that $|R| < 1$, (obviously necessary for (1) to hold) and $\hat{U} \sim N(U(1+R), \sigma^2)$, where σ^2 is known. Then \hat{U}^2/σ^2 has a non-central $\chi^2_{(1)}$ distribution with non-centrality parameter $\lambda = U^2(1+R)^2/2\sigma^2$. The null hypothesis $H_0: \sigma^2 \geq U^2(1-R^2)$ is equivalent to the hypothesis $H_0: \lambda \leq (1+R)/2(1-R)$. One approach, therefore, could be to adjust whenever $\hat{U}^2/\sigma^2 > c$, where the critical value $c \geq 0$ is chosen so that

$$\alpha = \Pr\left\{\chi^2_{\left(1, \frac{1+R}{2(1-R)}\right)} \geq c\right\}, \quad (2)$$

where α is the significance level of the test. This is a special case of a more general test suggested by Toro-Vizcarrondo and Wallace (1968).

Note that \hat{U}^2/σ^2 is the inverse of the square of the estimated coefficient of variation (CV) of \hat{U} . Thus, the criterion for adjustment can be interpreted in terms of a requirement to have a sufficiently small (in absolute value) CV.

In practice, we would have to substitute some prior estimate of the relative bias, say r , for R in (2). The sensitivity of c to various values of R is examined in Royce (1991) for the case of a one-sided test (a normal distribution was used instead of a χ^2 in this case). For example, with a significance level of 2.5%, it was found that the critical CV was only reduced from 33.8% to 27.1% even when the relative bias was as much as 50%.

If σ is not known but an estimate $\hat{\sigma}$ is available, then a similar test can still be constructed by assuming that

$$\frac{\nu \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(\nu)} \quad (3)$$

independent of \hat{U} . This leads to a test based on a non-central F distribution. Further details on the construction of such tests are given in Judge and Bock (1978).

In order to determine the MSE of the preliminary test estimator, we note that it can be written as $\hat{U}^P = I\hat{U}$ where

$$\begin{aligned} I &= 1 \quad \text{if} \quad \frac{\hat{U}^2}{\sigma^2} > c \\ &= 0 \quad \text{if} \quad \frac{\hat{U}^2}{\sigma^2} \leq c. \end{aligned} \quad (4)$$

When σ^2 is known, the MSE of this estimator can be shown to be (see, for example, Judge and Bock 1978, p. 72)

$$\begin{aligned} \text{MSE}(\hat{U}^P) &= \sigma^2 + U^2 R^2 + (2U^2(1+R) - \sigma^2) \Pr\{\chi^2_{(3, \lambda)} \leq c\} - \\ &\quad U^2(1+R)^2 \Pr\{\chi^2_{(5, \lambda)} \leq c\}. \end{aligned} \quad (5)$$

Note that as $c \rightarrow \infty$, *i.e.* as the chance of adjustment goes to zero, the MSE approaches U^2 , the MSE of the unadjusted census. Similarly, as $c \rightarrow 0$, *i.e.* as the chance of adjustment goes to certainty, the MSE approaches $\sigma^2 + U^2 R^2$, the MSE of the adjusted census estimator. Thus, the two previous approaches of adjustment or no adjustment can be interpreted as extreme cases of the pretest estimator procedure.

Figure 1 shows MSE/σ^2 for the preliminary test estimator as a function of U^2/σ^2 , for various values of c , in the unbiased case ($R = r = 0$). The MSEs/ σ^2 of the unadjusted census and the adjusted census are also shown. In all cases, the MSE of the preliminary test estimator starts out higher than that of the unadjusted census, crosses the MSE of the adjusted census, reaches a maximum, and then approaches the MSE of the adjusted census. As the value of c decreases and the level of significance α of the test therefore increases, the MSE of the preliminary test estimator approaches that of the adjusted census more quickly, but at the expense of being higher for small values of U^2/σ^2 . Thus, the performance of the preliminary test estimator over the range of possible values of U^2/σ^2 depends on the level of significance that is chosen for the test.

Figures 2 and 3 show similar plots in the case where $R = .5$ and $R = -.5$ respectively (since we may feel we have no information on which to base an estimate of R , we have set $r = 0$). Again, the MSEs of the preliminary test estimators approach those of the adjusted census as U^2/σ^2 increases. With a positive bias the MSE of the preliminary test estimator approaches the MSE of the adjusted census more quickly than in the unbiased case, while for a negative bias the reverse is true.

What is the “best” value of c for the test? Ideally, we would like to choose c so that the MSE of the preliminary test estimator is as close as possible to the minimum of the MSEs of the adjusted census and the unadjusted census. One approach, due to Sawa and Hiromatsu (1973) and extended by Brook (1976), is to minimize the maximum difference between the MSE of the preliminary test estimator and the minimum of the MSEs of the adjusted census and unadjusted census. For the unbiased case this criterion gives an optimal value of c of approximately 1.88. This corresponds to a critical CV (in absolute value) for the estimated under-coverage of 73%. The MSE of this estimator is shown in Figure 4.

Judge and Bock (1978) also describe other approaches to choosing the optimal value of c , such as minimizing the average distance (rather than the maximum difference) and Bayesian approaches.

2.4 Composite Estimator

The preliminary test estimator was written as $\hat{U}^P = I\hat{U}$, where I took on only the values 0 or 1. Because of this inherent discontinuity, however, it has been shown that the preliminary test estimator is inadmissible (Cohen 1965). As an alternative, we might consider letting the multiplier of \hat{U} take any value between 0 and 1. That is, instead of using the data to tell us **whether** to adjust, we use the data to tell us **how much** to adjust. This type of estimator has been suggested by Spencer (1980) and more recently by Andrews (1991). We define $\hat{U}^\alpha = \alpha\hat{U}$ where $0 \leq \alpha \leq 1$. For a given alpha, this estimator has a MSE equal to

$$\text{MSE}(\alpha\hat{U}) = \alpha^2\sigma^2 + U^2(\alpha(1 + R) - 1)^2, \quad (6)$$

which is minimized when

$$\alpha = \frac{U^2(1 + R)^2}{(1 + R)(\sigma^2 + U^2(1 + R)^2)}. \quad (7)$$

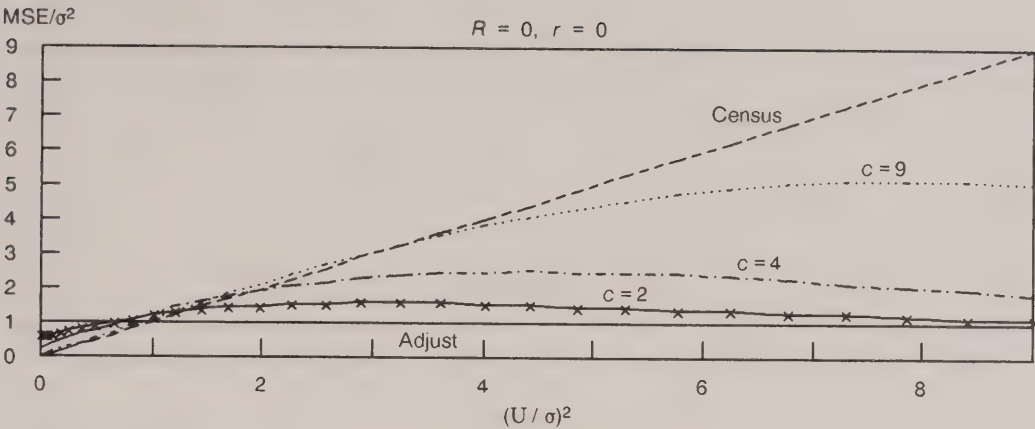


Figure 1 Comparison of MSEs

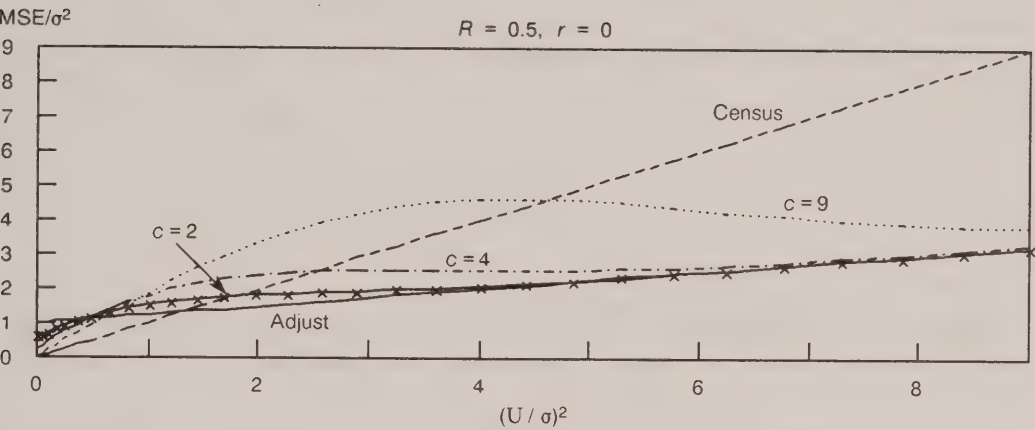


Figure 2 Comparison of MSEs

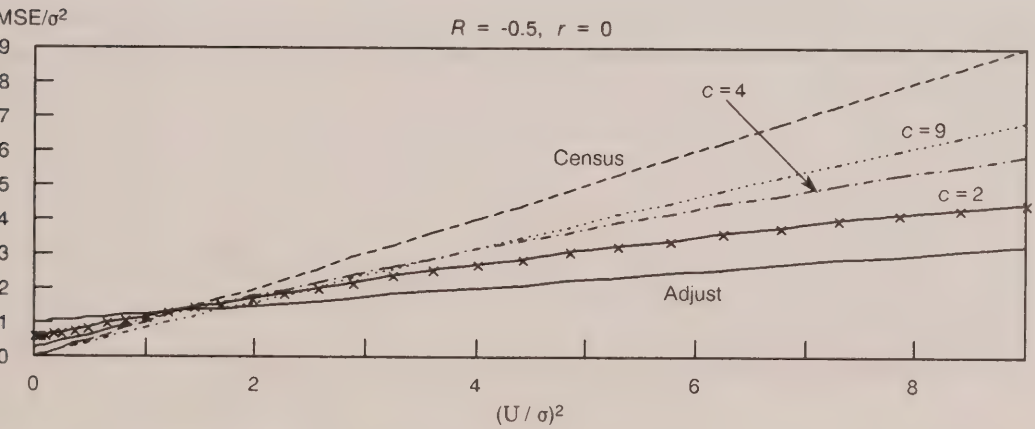


Figure 3 Comparison of MSEs

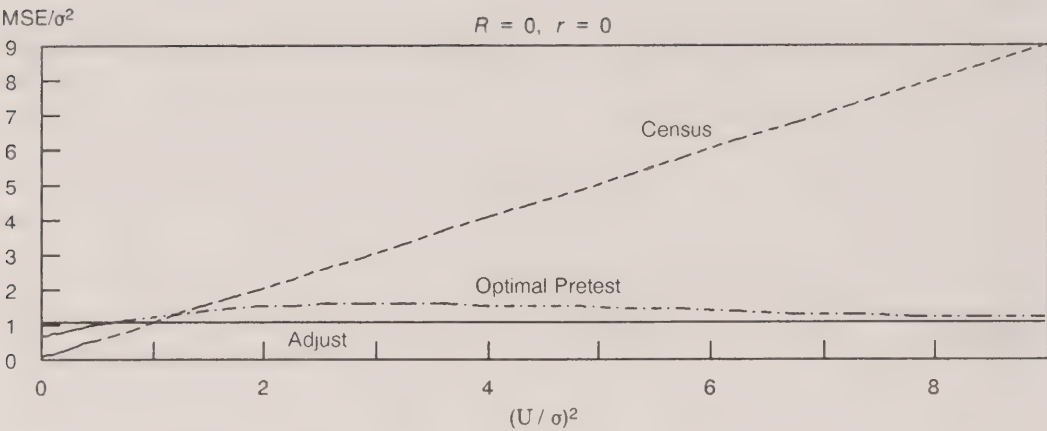


Figure 4 Comparison of MSEs

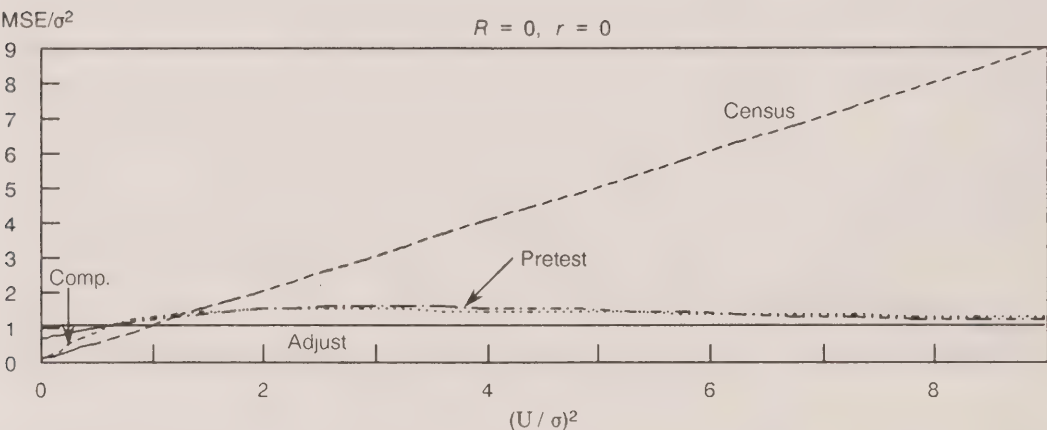


Figure 5 Comparison of MSEs

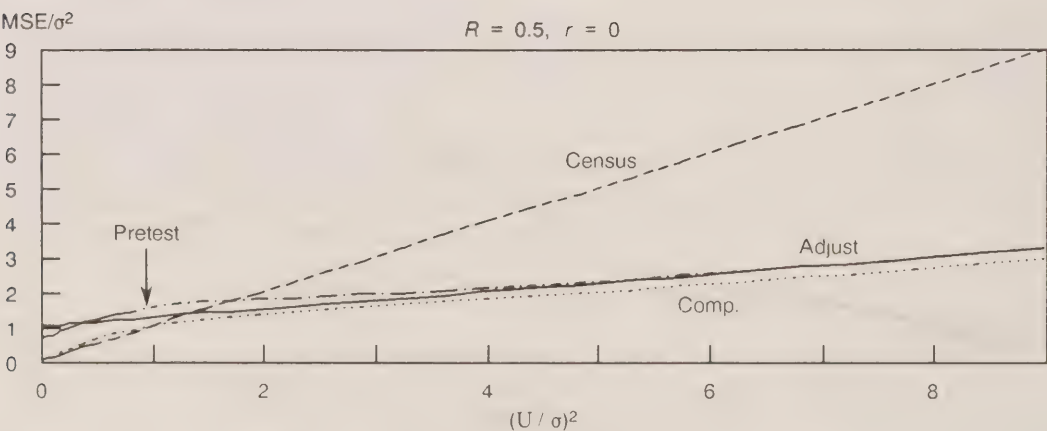


Figure 6 Comparison of MSEs

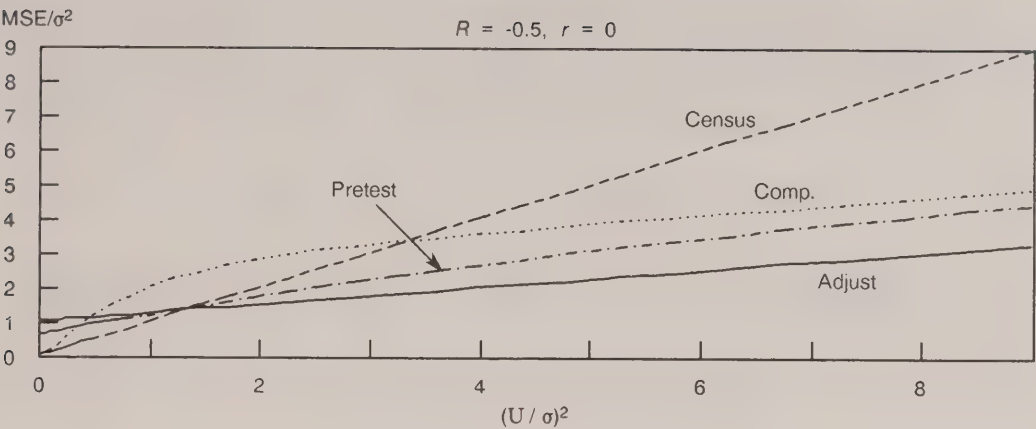


Figure 7 Comparison of MSEs

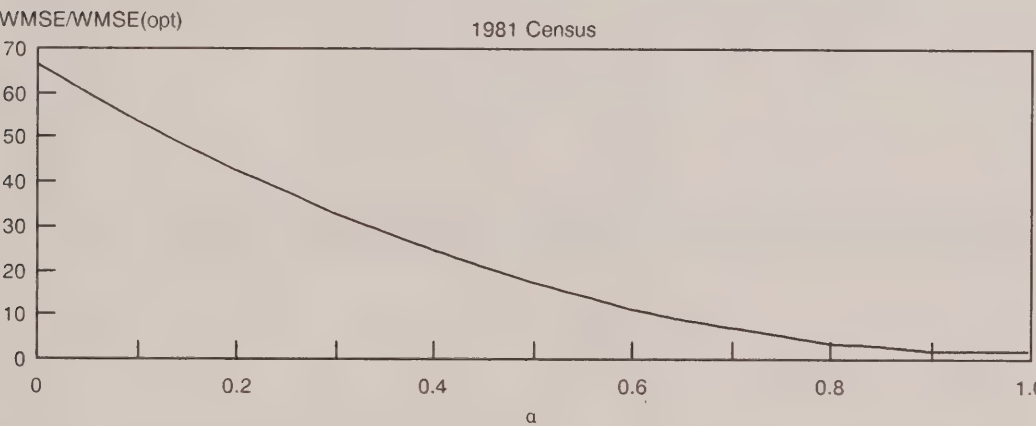


Figure 8 WMSEs for Totals

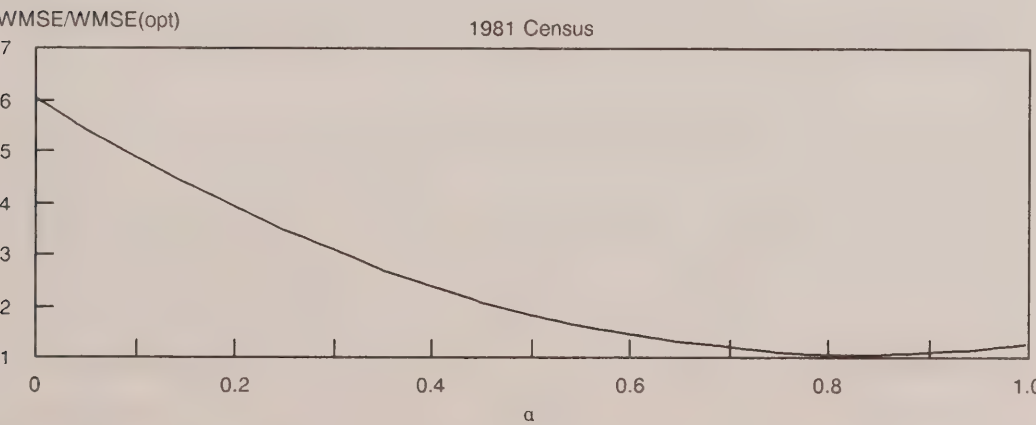


Figure 9 WMSEs for Shares

If σ is assumed known, then a possible estimator of α is

$$\hat{\alpha} = \frac{\hat{U}^2}{(1+r)(\sigma^2 + \hat{U}^2)} \quad (8)$$

and thus

$$\hat{U}^{\hat{\alpha}} = \frac{\hat{U}^3}{(1+r)(\sigma^2 + \hat{U}^2)}. \quad (9)$$

The approximate MSE of this estimator can be found using a Taylor series approximation. Letting

$$h(U, \sigma^2) = \frac{U^3}{(1+r)(\sigma^2 + U^2)} \quad (10)$$

we get (dropping terms higher than those involving the first derivative)

$$\begin{aligned} \text{MSE}(\hat{U}^{\hat{\alpha}}) &\doteq (h(U, \sigma^2) - U)^2 + \left(\frac{\partial h(U, \sigma^2)}{\partial U} \right)^2 (U^2 R^2 + \sigma^2) \\ &\quad + 2(h(U, \sigma^2) - U) \left(\frac{\partial h(U, \sigma^2)}{\partial U} \right) UR. \end{aligned} \quad (11)$$

This approximation can also be extended to the case where σ is unknown by making the assumption given in (3). The MSE is then increased by the additional term

$$\left(\frac{\partial h(U, \sigma^2)}{\partial \sigma^2} \right)^2 \frac{2\sigma^4}{\nu}. \quad (12)$$

Figures 5, 6 and 7 show the MSE of the composite estimator as a function of U^2/σ^2 , as well as the MSEs of the unadjusted census, adjusted census, and the optimal preliminary test estimator from Section 2.3. In the unbiased case (Figure 5) and the positive bias case (Figure 6), the composite estimator outperforms the optimal preliminary test estimator. When the bias is negative, however, (Figure 7) the MSE of the composite estimator can be much higher than any of the other estimators over a considerable portion of the range of U^2/σ^2 .

3. MORE GENERAL ESTIMATORS

In this section, we generalize the four estimators examined in Section 2 in two ways. First, instead of a single population total, we consider a vector of population totals, denoted as $\underline{T} = (T_1, T_2, \dots, T_N)$. Second, we consider not only the population totals themselves, but also functions of the population totals, denoted by $\underline{g}(\underline{T}) = (g_1(\underline{T}), g_2(\underline{T}), \dots, g_K(\underline{T}))$ where in general $K \neq N$. Typical functions of interest include population shares, used in the transfer of funds from the federal to provincial governments, as well as growth rates between censuses, differences in growth rates among different provinces, and so on.

In evaluating the overall accuracy of some estimate $\underline{g}(\hat{\underline{T}}^*)$ for $\underline{g}(\underline{T})$, we will make use of a loss function. The use of loss functions for evaluating the effects of census adjustment is

described in Fellegi (1980), Citro and Cohen (1985), Spencer (1986), and Wolter and Causey (1991) to name just a few. The specific loss function used in this paper is a generalization of previously-proposed loss functions for population totals and shares. Specifically, the risk (expected loss) of the estimator $\underline{g}(\hat{T}^*)$ is the Weighted Mean Square Error, defined as

$$WMSE(\underline{g}(\hat{T}^*)) = E \left\{ \sum_{k=1}^K w_k (g_k(\hat{T}^*) - g_k(T))^2 \right\}, \quad (13)$$

where w_k is a user-specified weight reflecting the importance of the k -th component of the loss function.

Since \underline{g} may be complex in practice, it is useful to work instead with an approximation to the WMSE derived by expanding $\underline{g}(\hat{T}^*)$ in a Taylor series around \underline{T} . This yields:

$$WMSE \underline{g}(\hat{T}^*) \doteq \sum_{i=1}^N \sum_{j=1}^N \omega_{ij} [\text{Cov}(\hat{U}_i^*, \hat{U}_j^*) + \text{Bias}(\hat{U}_i^*) \text{Bias}(\hat{U}_j^*)] \quad (14)$$

where the weight ω_{ij} is given by

$$\omega_{ij} = \sum_{k=1}^K w_k \frac{\partial g_k}{\partial T_i} \frac{\partial g_k}{\partial T_j}. \quad (15)$$

(Note that the approximate WMSE can also be written as the expectation of the quadratic form $(\hat{U}^* - \underline{U})' \Omega (\hat{U}^* - \underline{U})$ where ω_{ij} is the ij -th element of Ω .)

This formulation conveniently splits each component of the risk function into two parts: a weight ω_{ij} that depends only on the w_k and the function \underline{g} , and the portion in square brackets which depends only on the particular estimator being used.

While the choice of the w_k can be arbitrary, considerations of equity have often led to the choice $w_k = 1/T_k$. In the case of population totals and shares, for example, the risk function (14) then becomes equivalent to those proposed by Fellegi (1980) and also used by Wolter and Causey (1991), among others. Other choices for the weights that have been suggested in the literature include $w_k = 1/Y_k$, $w_k = 1/\hat{T}_k$, and $w_k = 1$. For further discussion on the merits of these various weightings, see the references cited above. Table 1 shows some examples of ω_{ij} for different functions.

In the case of population growth rates, the first pair of subscripts on the omega refer to the population quantity of interest (e.g. province) while the second pair refer to the census at time 1 or time 2 respectively. The second subscript on the T_i also refer to the census at time 1 or 2.

In the remainder of this section, we illustrate the use of the WMSE in developing and evaluating the unadjusted census, adjusted census, preliminary test estimator, and composite estimator.

3.1 Unadjusted Census

The WMSE of the unadjusted census is $WMSE(\underline{U}^C) = \sum_{ij} \omega_{ij} U_i U_j$.

3.2 Adjusted Census

The WMSE of the adjusted census is $WMSE(\underline{U}^A) = \sum_{ij} \omega_{ij} [\sigma_{ij} + b_i b_j]$ where $\sigma_{ij} = \text{Cov}(\hat{U}_i, \hat{U}_j)$ and $b_i = \text{Bias}(\hat{U}_i)$.

Table 1
Examples of Weights ω_{ij} in the Approximate WMSE for Various Functions

Function	ω_{ij}
Set of Population Totals	$\omega_{ii} = w_i$ $\omega_{ij} = 0 \quad i \neq j$
Set of Population Shares	$\omega_{ii} = \frac{1}{T^4} \left(\sum_k w_k T_k^2 + w_i T^2 - 2w_i T T_i \right)$ $\omega_{ij} = \frac{1}{T^4} \left(\sum_k w_k T_k^2 - T(w_i T_i + w_j T_j) \right) \quad i \neq j$
Set of Growth Rates	$\omega_{ii11} = \frac{w_i T_{i1}^2}{T_{i1}^4}$ $\omega_{ii12} = -\frac{w_i T_{i1} T_{i2}}{T_{i1}^4} = \omega_{ii21}$ $\omega_{ii22} = \frac{w_i T_{i1}^2}{T_{i1}^4}$ $\omega_{ij11} = \omega_{ij12} = \omega_{ij21} = \omega_{ij22} = 0 \quad i \neq j$

3.3 Preliminary Test Estimator

As in Section 2.3, we would use the adjusted census in preference to the unadjusted census if the WMSE of the adjusted census is less than the WMSE of the unadjusted census, *i.e.*, if

$$D = \sum_{ij} \omega_{ij} [U_i U_j - \sigma_{ij} - b_i b_j] > 0. \tag{16}$$

Tests for this type of hypothesis were suggested by Fellegi (1980) for the specific cases of population totals and population shares, but the ideas generalize quite readily to any function g . The left hand side of the inequality (16) is estimated by $\hat{D} = \sum_{ij} \omega_{ij} [\hat{U}_i \hat{U}_j - 2\sigma_{ij}]$ where the ω_{ij} are assumed to be known. (In practice the ω_{ij} are estimated by substituting either the census counts or the adjusted census counts in (13). Fellegi claimed that minor variations in the weights were unlikely to substantially change the test results.) It is then easy to show that $E(\hat{D}) = D + 2 \sum_{ij} \omega_{ij} b_i (U_j + b_j)$. For the case of totals and shares, Fellegi presented arguments why it could be assumed that the second term was non-positive, *i.e.*, $\sum_{ij} \omega_{ij} b_i (U_j + b_j) \leq 0$ so that \hat{D} would tend to underestimate D . Fellegi also derived an approximate variance for \hat{D} . This, along with the assumption that \hat{D} was normally distributed, permitted the construction of a test for the hypothesis given in (16).

Table 2
z Values for Fellegi's Tests for Adjustment of Provincial Population Totals and Shares, Reverse Record Check, 1976, 1981 and 1986

Function	1976	1981	1986
Totals	9.3	10.1	13.1
Shares	3.1	1.8	1.5

In the more general case, the approximate variance of \hat{D} is given by $\text{Var}(\hat{D}) \doteq 4 \sum_{ij} \sigma_{ij} (\sum_{i'j'} \omega_{ij'} \omega_{i'j} U_{i'} U_{j'})$. An estimate of $\text{Var}(\hat{D})$ can then be derived by substituting estimates of the U_i and σ_{ij} in this formula.

In the case of totals, for example, the test statistic (z value) is given by

$$z = \frac{\hat{D}}{\sqrt{\text{Var}(\hat{D})}} = \frac{\sum_i \frac{\hat{U}_i^2 - 2\sigma_i^2}{Y_i}}{2\sqrt{\sum_i \frac{\hat{U}_i^2 \hat{\sigma}_i^2}{Y_i^2}}}, \tag{17}$$

where in this case the inverse of the census counts have been used as the weights. A similar expression can be derived for population shares.

Table 2 shows the z values calculated for the censuses of 1976, 1981 and 1986 for provincial population totals and shares. The data come from the Reverse Record Checks conducted in these censuses.

The case for adjusting population totals is much stronger than the case for adjusting shares, reflecting the fact that estimates of differences in undercoverage rates among provinces are less accurate than estimates of the undercoverage rates themselves. Further numerical results are given in Royce and Luc (1990).

3.4 Composite Estimator

A natural extension of the composite estimator of Section 2.4 would at first seem to be $\alpha_i \hat{U}_i$. However the use of different amounts of adjustment for each value of i introduces problems of consistency. For example, it would imply that more adjustment should be done at the Canada level than at the province level, since the estimates of undercoverage at the province level will be less accurate than for the national level. If this were done, the provincial totals would not add up to the Canada total.

In practice, therefore, we constrain ourselves to a single value of alpha, *i.e.* $\underline{U}^\alpha = \alpha \underline{\hat{U}}$, where again $0 \leq \alpha \leq 1$. The WMSE of this estimator is

$$\text{WMSE}(\underline{U}^\alpha) = \sum_{ij} \omega_{ij} [\alpha^2 (\sigma_{ij} + b_i b_j) + (\alpha - 1)^2 U_i U_j + 2\alpha (\alpha - 1) U_i b_j], \tag{18}$$

which is minimized when

$$\alpha = \frac{\sum_{ij} \omega_{ij} U_i (U_j + b_j)}{\sum_{ij} \omega_{ij} [\sigma_{ij} + (U_i + b_i) (U_j + b_j)]}. \quad (19)$$

If, as was done in Section 3.3, we make the assumption that $\sum_{ij} \omega_{ij} b_i (U_j + b_j) \leq 0$ then a lower bound for the optimal alpha is given by

$$\alpha_L = \frac{\sum_{ij} \omega_{ij} (U_i + b_i) (U_j + b_j)}{\sum_{ij} \omega_{ij} [\sigma_{ij} + (U_i + b_i) (U_j + b_j)]}, \quad (20)$$

which we estimate by

$$\hat{\alpha}_L = \frac{\sum_{ij} \omega_{ij} \hat{U}_i \hat{U}_j}{\sum_{ij} \omega_{ij} [\hat{\sigma}_{ij} + \hat{U}_i \hat{U}_j]}, \quad (21)$$

assuming the ω_{ij} are known. In practice, as we did for the preliminary test estimator, we would estimate the ω_{ij} by substituting census counts or adjusted census counts in (15).

In the case of population totals, for example, the estimated amount of adjustment is

$$\hat{\alpha}_L = \frac{\sum_i \hat{T}_i \hat{U}_i^2}{\sum_i \hat{T}_i [\hat{\sigma}_i^2 + \hat{U}_i^2]}, \quad (22)$$

where \hat{U}_i is the estimated undercoverage rate, *i.e.* $\hat{U}_i / (Y_i + \hat{U}_i)$, and $\hat{\sigma}_i^2$ is its estimated variance.

For shares, the amount of adjustment is given by

$$\hat{\alpha}_L = \frac{\sum_i \hat{T}_i \hat{U}_i^2 - \hat{T} \hat{U}^2}{\sum_i \hat{T}_i [\hat{\sigma}_i^2 + \hat{U}_i^2] - \hat{T} (\hat{\sigma}^2 + \hat{U}^2)}, \quad (23)$$

where \hat{U} is the estimated undercoverage rate for the total population, *i.e.* $\sum_i \hat{U}_i / \sum_i (Y_i + \hat{U}_i)$ and $\hat{\sigma}^2$ is its estimated variance. The inverse of the adjusted census counts have been used as the weights in these two examples.

3.5 Numerical Comparisons

In the case of a single population total, it was possible to derive exact or approximate formulae for the MSEs of the four estimators as a function of U^2/σ^2 , R , r and (in the case of the preliminary test estimator), the critical value of the test. Unfortunately, it has not yet been possible to derive similar expressions for the WMSEs of complex functions of a vector of population totals.

In the case of the unadjusted census, adjusted census, and composite estimator, however, it is possible to estimate the WMSEs by substituting estimates of undercoverage and their estimated variances into equation (18) (if estimates of the bias terms are available they can be used, but in what follows we assume they are zero). For example, Figures 8 and 9 show, for the 1981 Census, the estimated ratio of the WMSE to the optimal WMSE, as a function of α , where the provinces are again the units indexed by i . The extremes of $\alpha = 0$ and $\alpha = 1$ correspond to the unadjusted and adjusted census counts respectively, while the minimum point on the curve corresponds to the optimal α . Figure 8 is for totals and Figure 9 is for shares. The optimum values of α were computed using formulae (22) and (23).

In each case, the optimal degree of adjustment is close to 1.0, and results in a WMSE considerably lower than the WMSE corresponding to no adjustment (*e.g.* by a factor of almost 70 for totals). The optimal degree of adjustment is less for shares than for totals, again reflecting the fact that estimates of differences in coverage rates between provinces are less accurate than the estimates of the rates themselves. It is also interesting to note that the WMSE for full adjustment is only slightly higher than that of the optimal degree of adjustment. This can have important practical significance, since it is much easier to explain a full adjustment to data users than to explain a partial adjustment.

4. SMALL AREA ESTIMATION

The previous two sections considered the case where direct estimates of undercoverage, and estimates of their variances, were available from the coverage measurement studies. This situation applies, for example, for provinces, for some major Census Metropolitan Areas, and for broad demographic groups (*e.g.* age by sex, age by marital status) at the national level. However the Population Estimates Program produces estimates at very detailed levels, such as single years of age by sex by marital status for some 260 Census Divisions. Direct estimates of undercoverage generally do not exist at such levels.

Nevertheless, the need to maintain consistency of the estimates requires that any adjustment made at a higher level be "carried down" to the detailed levels used by the estimates program. In this section, we consider the use of synthetic estimation for this purpose, and show how the WMSE can again be used to develop preliminary test estimators and composite estimators.

The synthetic estimator is based on the assumption that net undercoverage is uniform within each of a number of "adjustment groups", indexed by a . The synthetic estimate is then given by $\hat{U}_i^S = \sum_a \lambda_{ia} \hat{U}_a$ where $\lambda_{ia} = Y_{ia}/Y_a$. For example, the adjustment groups might correspond to age-sex groups, for which estimates of undercoverage \hat{U}_a are available at some higher level.

A special case of the synthetic estimator arises when there is only one adjustment group. Wolter and Causey (1991) have called this the across-the-board estimator. It is defined as $\hat{U}_i^{ATB} = \lambda_i \hat{U}$ where $\lambda_i = Y_i/Y$. WMSEs for the across-the-board and the synthetic estimator can be derived using equation (14). Since the ω_{ij} do not depend on the particular estimator used, only the portion in square brackets changes. Table 3 compares the estimators of U_i and their covariance and bias terms for the census, adjusted census, across-the-board and synthetic estimators.

Table 3

Examples of Covariance and Bias Terms in the Approximate WMSRE for Various Estimators

Estimator	\hat{U}_i^*	$\text{Cov}(\hat{U}_i^*, \hat{U}_j^*)$	$\text{Bias}(\hat{U}_i^*)$
Census	0	0	$-U_i$
Adjusted Census	\hat{U}_i	σ_{ij}	b_i
Across-the-Board	$\lambda_i \hat{U}$	$\lambda_i \lambda_j \sigma^2$	$\lambda_i (U + b) - U_i$
Synthetic	$\sum_a \lambda_{ia} \hat{U}_a$	$\sum_{aa'} \lambda_{ia} \lambda_{ja'} \sigma_{aa'}$	$\sum_a \lambda_{ia} (U_a + b_a) - U_i$

where $b = \sum_i b_i$ and similarly b_a is the bias of \hat{U}_a .

4.1 Preliminary Test Estimators

As was the case in Sections 2 and 3, the WMSE can be used to develop statistical tests to decide between two competing estimators. As an example, consider the situation where we wish to choose between the unadjusted census and the across-the-board estimator for population totals (shares are of course unchanged by across-the-board adjustment). On comparing the WMSEs of these two estimators, we find that we would use the across-the-board estimator in preference to the census counts if

$$\sigma^2 < U^2(1 - R^2) \left[1 - \frac{2TB}{U(1 - R)} \right], \tag{24}$$

where

$$B = 1 - \frac{1}{\sum_i \frac{\lambda_i^2}{\tau_i}}, \tag{25}$$

and $\tau_i = T_i/T$. This condition was given, in a different form, by Wolter and Causey (1991). B is a measure of the heterogeneity of undercoverage; it is non-negative, and is equal to zero if and only if the undercoverage is completely uniform.

Noting that this inequality is the same as (1) except for the additional term in square brackets, we can derive a test very similar to the test described in Section 2.3. The critical value of the coefficient of variation will depend on the chosen significance level and the relative bias as before, but will also depend on B/\bar{U} , the ratio of the heterogeneity of undercoverage to the overall undercoverage rate.

Royce (1991) showed that, in practice, the effect of this additional factor on the critical CV was likely to be negligible. Thus, if adjustment is justified at some higher level, then carrying down the adjustment to lower levels is almost certainly justified as well. Similar results were found in a simulation study reported by Wolter and Causey (1991).

4.2 Composite Estimators

In Sections 2 and 3 we considered composite estimators where the two extremes were the unadjusted census and the adjusted census. With the addition of the synthetic and across-the-board estimators, the number of possible composite estimators increases considerably. For example, we might consider composite estimators involving the unadjusted census and the synthetic estimator, the adjusted census and the across-the-board estimator, the across-the-board estimator and the synthetic estimator, and so on. Consequently, we present below a method which can be used to derive a composite estimator involving any two estimators.

Our general composite estimator is defined as $\hat{U}^* = \alpha \hat{U}_1 + (1 - \alpha) \hat{U}_2$ where \hat{U}_1 and \hat{U}_2 are two estimators. The WMSE of this estimator is

$$\begin{aligned} \text{WMSE}(g(\hat{T}^*)) &= \alpha^2 \text{WMSE}(g(\hat{T}_1)) + (1 - \alpha)^2 \text{WMSE}(g(\hat{T}_2)) \\ &\quad + 2\alpha(1 - \alpha) \text{WMXPE}(g(\hat{T}_1), g(\hat{T}_2)), \end{aligned} \quad (26)$$

where

$$\text{WMXPE}(g(\hat{T}_1), g(\hat{T}_2)) = \sum_{ij} \omega_{ij} [\text{Cov}(\hat{U}_{1i}, \hat{U}_{2j}) + \text{Bias}(\hat{U}_{1i}) \text{Bias}(\hat{U}_{2j})] \quad (27)$$

is defined to be the Weighted Mean Cross-Product Error of $g(\hat{T}_1)$ and $g(\hat{T}_2)$. The WMSE of our composite estimator is minimized when

$$\alpha = \frac{\text{WMSE}(g(\hat{T}_2)) - \text{WMXPE}(g(\hat{T}_1), g(\hat{T}_2))}{\text{WMSE}(g(\hat{T}_1)) + \text{WMSE}(g(\hat{T}_2)) - 2\text{WMXPE}(g(\hat{T}_1), g(\hat{T}_2))}. \quad (28)$$

To obtain an estimate of α , we substitute estimates of the WMSEs and the WMXPE into the above.

As an example of how this approach could be used, suppose a decision has been taken to adjust a provincial population total. To carry down the adjustment, we might consider using either across-the-board adjustment (*i.e.* adjust all sub-provincial quantities by the same factor), or a synthetic adjustment, where the adjustment is done separately within several age-sex groups. The across-the-board method has the advantage that it uses only the provincial estimate of undercoverage, which is likely to be more reliable than the estimates of undercoverage by age and sex at the province level. On the other hand, if undercoverage varies considerably among age and sex groups, and if the sub-provincial quantities indexed by i also differ in their age-sex composition, then the synthetic estimator may be better.

If estimates of the U_i are available from some source, then all covariance and bias components of the WMSEs and the WMXPE can be estimated (using formulae such as those in Table 3), and the optimum composite estimator involving the across the board and synthetic estimators can be estimated. Although for sub-provincial quantities the U_i will not usually exist in practice, the method can be investigated at higher levels. For example we could use the provinces as the quantities indexed by i and use across-the-board and synthetic adjustment factors computed at the Canada level. A second possibility is to construct an artificial population (*e.g.* as in Shirm and Preston (1987) or Wolter and Causey (1991)) where the U_i are assumed to be known.

5. FURTHER WORK

The results presented in this paper represent only a start to the investigation and comparison of the performance of various estimators of a set of population totals. There are several areas where considerable work is yet required.

First, further investigation of the WMSEs for the preliminary test and composite estimators in the more general cases described in Sections 3 and 4 is required. Although attempts to derive analytic expressions for these WMSEs have not yet been successful, the more general results for preliminary test estimators and Stein-rule estimators described by Judge and Bock (1978) may yet be found to apply. If so, this would help to answer questions such as: Can optimal critical values be found for the Fellegi-type preliminary test estimators of Sections 3.3 and 4.1? How does the WMSRE of the preliminary test estimator compare in practice to those of the other three estimators?

Second, more work is needed to explore the sensitivity of the results to different weightings in the loss function. The results of Section 3 were based on the use of a weight equal to the inverse of the census count or the adjusted census count for each province. If the provinces had been weighted differently, the results would change. A more general weight we might want to consider is $w_k = Y_k^\gamma$, where γ is some type of power parameter. The sensitivity of the results in Section 3 to various values of γ could then be studied.

Finally, while the methods described in this paper provide a framework for developing and evaluating various estimators, the exact manner in which the methods will be applied has yet to be decided. Specific issues that must be resolved include:

1. What is the relative importance of different types of functions such as totals, shares and growth rates? Different functions give rise to different results, but in the end a single estimator must be chosen in order to maintain consistency.
2. At what geographic and demographic levels should these methods be applied? For example, should the preliminary test estimator or composite estimator described in Section 3 be applied at the province level, at the province by age group and sex level, or at even more detailed levels? The results obtained depend on the level of analysis used.
3. Could we even consider composite estimators for “high profile” estimators such as the provincial population totals? It might be difficult to explain to users why the adjustments do not coincide with the published estimates of undercoverage.

Because the resolution of issues such as these will require professional judgement, the decision about whether to adjust (and how to adjust) cannot be an automatic one based on completely pre-specified criteria. While the methods described in this paper can provide useful guidance, the final decision will require a careful balancing of the potential improvement in the accuracy of the estimates with consideration of how easily the methods can be communicated to and understood by users of the estimates program.

ACKNOWLEDGEMENTS

I would like to thank the editor, the two referees, and Richard Carter for several valuable comments that helped improve the quality of this paper.

REFERENCES

- ANDREWS, D. (1991) Invited discussion at the Meeting of Statistics Canada's Advisory Committee on Statistical Methods, October 1991.
- BROOK, R.J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association*, 71, 353, 126-131.
- CITRO, C.F., and COHEN, M.L. (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.
- COHEN, A. (1965). Estimates of the linear combination of parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics*, 36, 78-87.
- FELLEGI, I.P. (1980). Should the census count be adjusted for allocation purposes? Equity considerations. In *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census.
- JUDGE, G.G., and BOCK, M.E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam-New York-Oxford: North-Holland Publishing Company.
- ROYCE, D., and LUC, M. (1990). Recalculation of Fellegi's test statistics on census adjustment for the 1981 and 1986 censuses. Internal Statistics Canada report.
- ROYCE, D. (1991). Technical criteria for adjusting the population estimates program for census coverage error. Internal Statistics Canada report.
- ROYCE, D. (1992). Incorporating estimates of census coverage error into the Canadian population estimates program. *Proceedings of the Eighth Annual Research Conference*, Bureau of the Census, Washington, DC (to appear).
- SAWA, T., and HIROMATSU, T. (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica*, 41, 1093-1101.
- SHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 400, 965-978.
- SPENCER, B. (1980). Implications of equity and accuracy for undercount adjustment: A decision-theoretic approach. In *Proceedings of the 1980 Conference on Census Undercount*, Bureau of the Census, Washington, DC.
- SPENCER, B. (1986). Conceptual issues in measuring improvement in population estimates. In *Proceedings of the Second Annual Research Conference*, Bureau of the Census, Washington, DC, 393-407.
- TORO-VIZCORRONGO, C., and WALLACE, T.D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 322, 558-572.
- WOLTER, K.M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 414, 278-284.

The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census

L. SWAIN, J.D. DREW, B. LAFRANCE and K. LANCE¹

ABSTRACT

The Address Register is a frame of residential addresses for medium and large urban centres covered by Geography Division's Area Master File (AMF) at Statistics Canada. For British Columbia, the Address Register was extended to include smaller urban population centres as well as some rural areas. The paper provides an historical overview of the project, its objective as a means of reducing undercoverage in the 1991 Census of Canada, its sources and product, the methodology required for its initial production, the proposed post-censal evaluation and prospects for the future.

KEY WORDS: Address Register; Census undercoverage; Geographical Information Systems (GIS).

1. OBJECTIVE

The concept of an Address Register at Statistics Canada dates back to the 1960s. Fellegi and Krótki (1967) first considered building one for the 1971 Census using administrative source files as the base. Their approach was mostly manual and yielded a very complete set of addresses with minimal undercoverage and overcoverage. In the mid-1970s (Booth 1976), the idea resurfaced in planning for the 1981 Census. This time the approach started with data capture of addresses from the previous Census and was augmented with information from Canada Post. In both cases, the generated address lists were being considered as a frame for a mail-out Census. However, costs of creation were high and would have needed offsetting reductions in other Census operations to be effective. In addition, the risks associated with changing the traditional enumeration method were considered too great. As a result, the construction of an Address Register was suspended in each case.

A renewed interest in the concept of an Address Register emerged from the International 1991 Census Planning Conference (Royce 1986, 1987) in October 1985. This interest derived from the potential for automation of Fellegi and Krótki's approach due to technological developments, such as the availability of machine readable administrative files with addresses and postal codes and the development of in-house software to parse addresses into standard components, to assign postal codes and to link postal codes to Census geography. It followed as well from the development of a statistical theory for record linkage (Fellegi and Sunter 1969) and computer systems based on this theory (Hill and Pring-Mill 1985).

As a result, a project was initiated in 1986 with the first research (Gamache-O'Leary *et al.* 1987) investigating the use of an Address Register for a mail-out Census rather than the traditional drop-off approach. It concluded that the new Census data collection approach would be less expensive only if the quality of the Address Register required minimal field updating prior to the Census. Two small pilot registers created in early 1987 put Address Register coverage at 90-95%, which was unacceptable without field updating (Drew *et al.* 1987), ruling out the use of an Address Register for a mail-out Census.

¹ L. Swain and B. Lafrance, Social Survey Methods Division; J.D. Drew, Household Surveys Division; K. Lance, Geography Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

However, the two pilot registers revealed the potential for an Address Register to aid in coverage improvement when used in conjunction with the traditional drop-off methodology. This fitted well with the emergence of coverage improvement as one of the top priorities for the 1991 Census. The results of the Reverse Record Check for the 1986 Census had indicated a dramatic rise in the undercoverage rate compared to previous Censuses (from 2.01% in 1981 to 3.21% in 1986 for the national total population; from 2.08% in 1981 to 3.28% in 1986 for the national urban population) (Statistics Canada 1990). It was therefore decided that *the research project should concentrate on the development of the Address Register to use in coverage improvement of the 1991 Census.*

The next section describes the two major tests conducted to develop and refine the procedures used to create the Address Register for the 1991 Census. As well, the second section outlines the joint agreement with the Province of British Columbia to extend the Address Register. The third section presents the administrative and geographic sources used in the production process and the structure and content of the Address Register booklets, the end product used by Census Representatives in the field. The fourth section describes the methodology used to exploit the sources in order to produce the Address Register booklets. In the fifth section, the proposed post-censal evaluation is discussed while the last section presents future prospects for the Address Register. A separate future report will detail an evaluation of the methodology.

2. BACKGROUND

2.1 The November 1987 Test of Coverage Improvement Methods

A substantial test of the use of the Address Register (AR) as a coverage improvement tool was conducted in November 1987 in five large Regional Office cities. It was designed to estimate both undercoverage and overcoverage of dwelling units for the traditional Census method of listing and for two experimental methods of using an AR for Census coverage improvement: Post-list and Pre-list. The Post-list approach had the enumerator compile the dwelling list in the usual Census manner (creating a Visitation Record) then reconcile it with a dwelling list for the Enumeration Area (EA) derived from the AR. Field follow-ups were done where necessary on any address discrepancies between lists. In the Pre-list method, the enumerator was given the AR in advance and updated it during a canvass of the EA to create the final dwelling list.

The results (van Baaren 1988) concluded that the Post-list method was the more effective in improving coverage. This approach as a simple add-on to the standard Census enumeration process was fail-safe. If for some reason we failed to produce the AR (either in whole or in part) on time for the 1991 Census, the AR reconciliation step could simply be dropped without affecting the traditional enumeration process. The test data also provided estimates of the degree of coverage improvement and costs (Royce and Drew 1988). It was estimated that 34,000 occupied dwellings and 68,000 persons would be added by the AR to the medium and large urban centres for which it would be constructed (these urban centres representing those areas for which an Area Master File exists, *i.e.*, covering about 65% of the Canadian population). This would represent an improvement in coverage of 0.26 percentage points (the national undercoverage rate in 1986 being estimated as 3.21 percent). Relative to the two previous attempts at AR construction, costs were demonstrated to be low to the Census due to the highly automated approach and the proven benefit. As well, the risk was minimized since the traditional data collection method would still be used. Based on this cost, benefit and risk assessment, approval was given for creation of an AR for the 1991 Census.

From the November 1987 test, two concerns presented themselves. First, the ordering of the addresses in the AR booklets produced for each Enumeration Area (EA) didn't correspond to the order in the Visitation Records which made reconciliation a tedious and time-consuming task. Second, the overall overcoverage at 17% still seemed too high and more effort was required to eliminate erroneously placed or duplicate records. Both these problems were addressed by improving the methods for matching the AR to Census geography. Instead of linking addresses merely to EAs as had been done for the November test, procedures were developed to match the AR to the Area Master File (AMF) (Statistics Canada 1988) blockfaces. An algorithm was developed to sort addresses by block and within block in the same order they would be encountered by the enumerator in walking around the EA.

2.2 The September 1989 Test to Refine Procedures

Another substantial test was conducted in September 1989 involving four cities of various sizes: Moncton, Laval, Brampton and Calgary. Each was chosen because of unique difficulties that could arise based on the November 1987 test. The results (Dick 1990) showed a significant decrease in coverage from 84% in the 1987 test to 73%, a discouraging outcome. On the other hand, this test revealed a considerable reduction in overcoverage down from 17% to 8%. Importantly, despite the reduced coverage of the AR, its performance as a coverage improvement tool for the Census was still viable. On analysis, the new geocoding operation was found to be problematic, both in terms of its high costs, since it involved a great deal of clerical intervention, and in terms of its quality. The geocoding steps were therefore revamped for production, a key aspect of which was the adoption of CANLINK record linkage software (Statistics Canada 1989b) to improve quality and reduce costs of the AR/AMF linkage.

2.3 Joint Agreement with the Province of British Columbia

The Ministry of Finance and Corporate Relations in British Columbia was concerned about the high rate of undercoverage in their province in the 1986 Census (4.49% in 1986, up from 3.16% in 1981, for the provincial total population) (Statistics Canada 1990). Statistics Canada entered into a joint agreement with the Planning and Statistics Division (the provincial statistical agency) of the Ministry to help reduce undercoverage in British Columbia in the 1991 Census. Within this contract, the Address Register was expanded to include smaller urban areas in British Columbia, thereby increasing the population covered from 62% to 88%.

3. SOURCES AND PRODUCT

Production started in April 1990 and ended with the final Address Register (AR) booklet stapled in mid-May 1991, when 22,756 booklets had been compiled containing 6.6 million addresses for use in the Census data collection process.

3.1 Administrative Sources

In the September 1989 test, it was concluded that wherever possible the following four administrative sources ought to be used as sources of addresses to create the AR: telephone company billing files, municipal assessment rolls, hydro company billing files and the T1 Personal Income Tax file. However, the use of all four sources was possible only in Nova Scotia, New Brunswick, and eight major urban centres in Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton and Windsor). Because of the multiplicity of files,

the cost of files and refusals, only three sources were used for Newfoundland, Québec, Manitoba, Alberta (telephone, hydro and tax files) and for Regina and the rest of Ontario (telephone, assessment and tax files). For Saskatoon, only telephone and tax files were available. The primary source files used by the British Columbia government were those of telephone and hydro, though motor vehicles, cable and Elections files were also used.

3.2 Geography Sources

In building the AR, extensive use was made of a Geography Division system and files.

- i. The Area Master File (AMF) (Statistics Canada 1988) is a digitized feature network (covering streets, railroads, rivers, *etc.*) for medium and large urban areas, generally with populations of 50,000 or more. Of interest for the AR were the street features which contained street name and civic number ranges which could be used to locate individual addresses onto a blockface, the primary linkage.
- ii. The Computer Assisted Mapping System (CAM) orders blockfaces into blocks and blocks into a Census Enumeration Area (EA). CAM was used for the sequencing of addresses in the AR booklets. The EA maps produced by CAM were used by the Census Representatives for the 1991 Census. For the AR, the maps for all AMF areas were used in the second clerical operation.
- iii. The 1990 Postal Code Conversion File (PCCF) (Statistics Canada 1991) is a national file of all postal codes, each of which is linked to a 1986 Census EA or a series of 1986 EAs. This input was used for secondary linkage of addresses at the EA level.
- iv. The 1986/1991 EA Correspondence File relates the 1986 EA geography to the 1991 geography. This file was used for the secondary linkage at the EA level and the second clerical operation.

3.3 Address Register Booklets

The end product consisted of a set of booklets of residential addresses, one for each Enumeration Area, covering urban areas of Canada for which an Area Master File existed. Figure 1 contains a fictitious example of a page from an AR booklet (reduced in size).

Each booklet was divided into two sections: a structured portion and an unstructured portion. The structured portion contained all the addresses tied to a blockface with all the blockfaces being sequenced into blocks within the EA. The sequencing mirrored that found on the map that the Census Representative (CR) used for listing the EA in his/her Visitation Record (VR). The unstructured portion contained the addresses that could be tied only to the EA rather than a blockface. These were sorted by odd/even civic numbers within street name. The volume of addresses split 90%-10% between structured and unstructured.

Besides the address data, each page in an AR booklet contained a series of columns to be used in the reconciliation operation between the AR and VR. In the reconciliation, the Census Representative manually compared the Visitation Record with the AR to identify matches and non-matches. If the address was only on the VR, it was added to the AR (undercoverage in the AR). If the address was only on the AR, field resolution was usually required by the CR, with the result that the address was designated either as a new address to be enumerated for the Census by the CR (undercoverage in the Census) or as an invalid address classified by type of error (overcoverage in the AR). Addresses were denoted as invalid if they were duplicates, if they lay outside the EA, or for any other reason. All valid addresses had the Census Household Number coded in the booklet by the CR. A telephone number for the address, if available,

ADDRESS REGISTER

Protected

PROVINCE 35
FED 038EA 261
VN 0Page 21 of 22

Block No.	Address			Hhld No.	Not Listed at Drop-off	Field Follow-up Required	Invalid			AR Ref No.	Telephone Number
	Civic No.	Street	Apt. No.				Duplicate	Outside EA	Other		
1	2	3	4	5	6	7	8	9	10	11	12
4	23	MAIN	ST							1044566	5551111
4	19	MAIN	ST							1044564	5561234
4	15	MAIN	ST							1044562	5552321
4	11	MAIN	ST							1044559	
4	9	MAIN	ST							1044583	7475739
4	7	MAIN	ST							1044581	5552222
5	30	CENTRE	RD							1019615	5561029
5	34	CENTRE	RD							1019617	
5	34	CENTRE	RD	BT						1019618	5564261
5	60	CENTRE	RD							1019627	
5	64	CENTRE	RD							1019629	7478765
5	68	CENTRE	RD							1019634	5556942
5	72	CENTRE	RD							1019636	
5	76	CENTRE	RD							1019640	
5	80	CENTRE	RD							1019642	7476789
5	84	CENTRE	RD							1019644	5568765
5	88	CENTRE	RD							1019646	5559999
5	92	CENTRE	RD							1019579	7473456
5	96	CENTRE	RD							1019581	7450987
5	100	CENTRE	RD							1019648	
5	108	CENTRE	RD							1019579	5557171
5	112	CENTRE	RD							1019581	5558888
5	116	CENTRE	RD							1019583	7462009
5	120	CENTRE	RD							1019586	7450235
5	124	CENTRE	RD							1019588	5569630

Figure 1. Example of a Page from an AR Booklet (reduced in size).

was pre-printed in the last column of the booklet to assist the CR in any required Census follow-up operation.

4. METHODOLOGY

In this section, the creation of the Address Register (AR) is described. Figure 2 provides an overview of the steps involved.

4.1 Overview of the Methodology

The free-format addresses contained on the source files were first standardized into ordered component parts (steps 1 and 2) in preparation for the use of subsequent software. Then, postal codes were confirmed or corrected (step 3) so that those areas or worksites for which the AR was to be created could be selected from among all the addresses and locations contained on the source files (step 4). Because the same addresses could be contained on more than one file or more than once on the same file, unduplication of addresses based on both exact and probabilistic matching took place (steps 5 and 6).

Next, automated linkages were made of addresses to the blockface level using the Area Master File (step 7) or, where this was not possible, to Enumeration Area (EA) using the Postal Code Conversion File (step 8). After loading the addresses into a database management system (step 9), manual linkages were made of addresses to blockface (steps 10 and 11) or to EA

(step 12). Addresses within each EA were then sequenced by and within blocks (step 13) before being printed and collated in booklets by EA (step 14) for use in the Census.

4.2 Address Standardization (Steps 1, 2 and 3)

The Postal Address Analysis System (PAAS – step 2 of Figure 2) (Statistics Canada 1989c) performed two tasks: it broke up the free-format addresses from the source files into their component parts (street name, civic number, street designator, street direction, apartment number, municipality, province, postal code) and composed the address search key (ASK). ASK is an ordered concatenation of all the components of an address and is used during unduplication.

Although PAAS was an excellent product, analysis from the 1989 prototype had revealed certain shortcomings that we felt could be resolved by grooming or filtering the administrative file contents prior to using the generalized software. This FILTER step (step 1) concentrated on the following tasks: eliminating special characters with which PAAS refused to deal, repackaging address components in a manner compatible with PAAS, translating street designator short forms to acceptable ones, introducing commas between the street and municipality components of the free-format address to improve PAAS's comprehension, eliminating leading zeroes from civic numbers and numeric street names, and adding municipality and province names.

The FILTER and PAAS steps were applied in an iterative fashion. The first step was to discover what anomalies needed filtering for each administrative source. If the PAAS error rate after filtering was greater than 5%, error records were reviewed to find recurring problems that could be successively eliminated by further filtering until an error rate of less than 5% was achieved. As any address record that failed address standardization was eliminated from further consideration, it was vital to have a PAAS success rate as high as possible.

The PCVERIFY step (step 3) used the Automated Postal Coding System (PCODE) (Statistics Canada 1989a) package for confirmation and generation of postal codes. It was not quite as effective as the PAAS software at address analysis and could only confirm or add postal codes for 84% of the output from PAAS. It confirmed 78% of the postal codes and changed another 6%. Only .003% of the source administrative records had arrived with no existing postal code. It was crucial to have correct postal codes because these would be used for worksite selection in the subsequent step.

Two problems arose in the PCVERIFY step during production. If an address was missing a municipality/province component, the software continued to attempt to find a postal code instead of suspending further processing. As a consequence, enormous amounts of processing time could be spent trying to find postal codes. This problem was solved by including in the FILTER a step to add municipality and province names. The second problem occurred when a street name was numeric, as the processing time per address increased fourfold. This problem was not resolved and will necessitate modifications to the PCODE software.

4.3 Worksite Selection (Step 4)

This step partitioned the country by postal code into manageable worksites for processing with the sizes of worksites being based on the efficiency of CANLINK software for linkage of multiple large files. A geographic partitioning into worksites was adopted so they had dwelling counts in the 100,000 to 150,000 range based on the 1986 Census. Worksites were formed from an individual AMF (for a medium sized city), collections of physically adjacent AMFs (for small towns/townships), or parts of an AMF (for a large city). Geography Division's

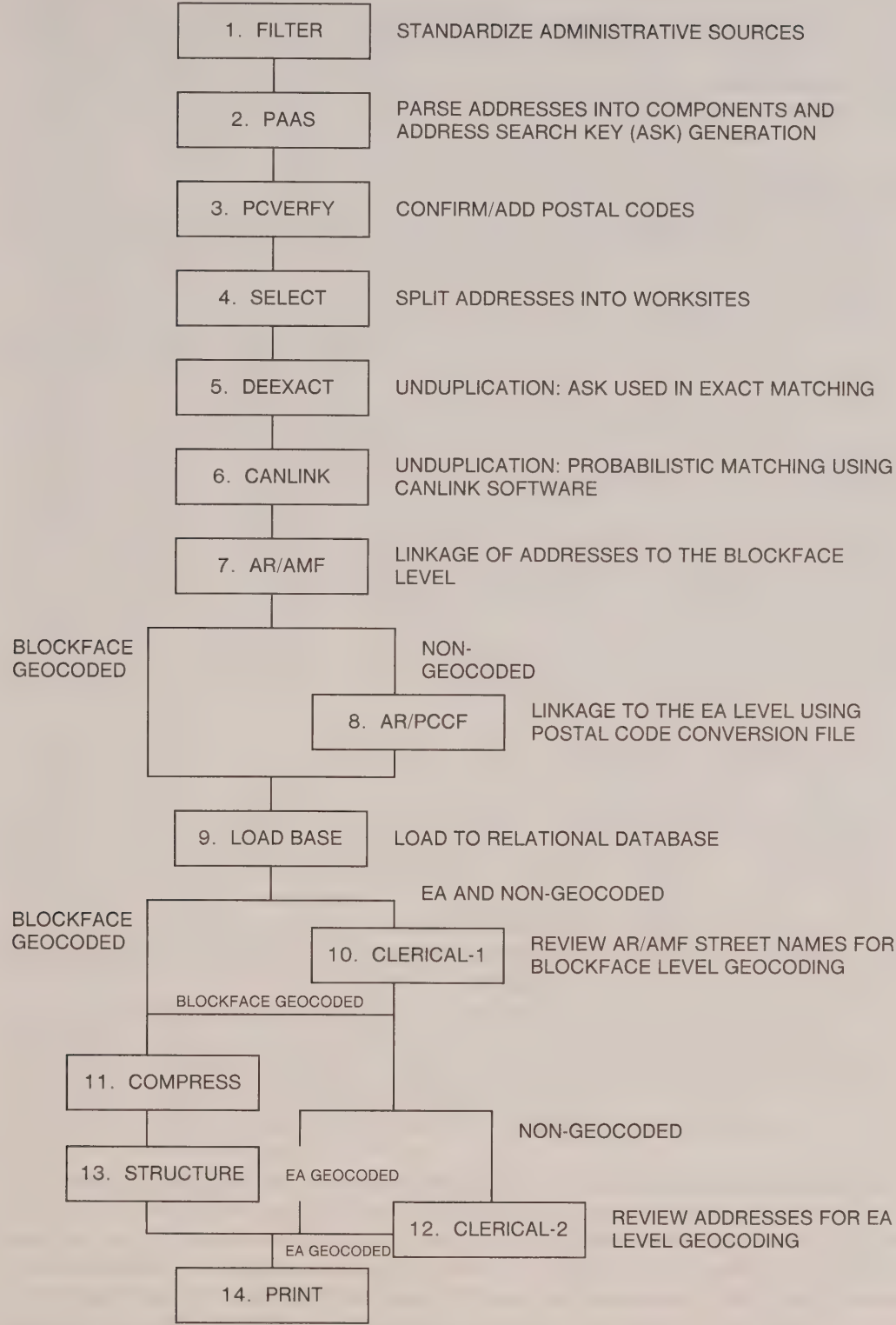


Figure 2. Overview of the Methodology.

Postal Code Conversion File (PCCF) which links postal codes and detailed Census geography was used to do this partitioning in the SELECT step (step 4). Once partitioning was completed, there were 105 distinct worksites and the original 43.4 million addresses had been reduced to 20.5 million addresses, with the dropped addresses having postal codes outside the AMF areas (*i.e.*, smaller cities and rural areas).

4.4 Unduplication (Steps 5 and 6)

In order to delete addresses included more than once on the source files, an unduplication process was conducted in two stages: an exact match with DEEXACT (step 5) and a probabilistic match using CANLINK software (step 6).

The DEEXACT step utilized the address search key (ASK) produced by the PAAS software and all records with an identical ASK were collapsed into a single record. With DEEXACT, the 20.5 million records from the SELECT step were reduced down to 10.1 million records. This reduction shows the importance of performing the address standardization.

Step 6 utilized the CANLINK generalized record linkage software (Statistics Canada 1989b). It clusters close records into groups called “pockets” and only records within the same pocket are actually matched together. For this application, civic number was used as the pocket. The components of the address (street name, municipality name, postal code, *etc.*) were used for matching purposes and weights were assigned for agreement or disagreement of each component. The development of levels of partial agreement for street name, municipality name and the last three characters of the postal code allowed for spelling variations and letter transpositions within the fields. The CANLINK step accounted for a further reduction to 6.7 million records. More details on the use of CANLINK in address unduplication are given in Drew *et al.* (1988), where its application in the November 1987 test is described.

4.5 AR/AMF Linkage (Step 7)

The major concern from the 1989 test was the strategy used to link addresses to their respective blockface. Because of the 11% drop in coverage from 84% to 73% compared to the 1987 test, a thorough investigation was needed and possibly a new approach. The other concern was that automated matching accounted for only 80% of the records matched while the other 20% were picked up clerically. This would have represented a daunting manual workload in full production. In order to circumvent these two concerns, another CANLINK application was developed for the AR/AMF linkage (step 7).

The original 1989 test files for Brampton still existed, so this became the test site for developing this step. The revised approach yielded 10% more matches, which increased the coverage back up to 1987 levels. As well, the automated matching was now responsible for 97% of the matches with 3% being picked up clerically, a significant improvement on the earlier 80%-20% split. Based on these results, the CANLINK approach was adopted for Census production.

In the construction of the new matching strategy, the first area of study involved a comparison of the contents of fields that would be used for matching purposes. This revealed certain anomalies that could be corrected prior to use to improve the number of linkages. The processing modifications to existing fields covered the following areas: removal of blanks between compound street names; alignment of street directions and civic numbers; conversion of numeric street names to numbers (on the AMF); removal of special characters in street names (on the AMF); correction of spelling variations in municipalities (on the AR); and a

recreation of certain PAAS translations for street names (on the AR). Several new fields were also generated: NYSIIS (New York State Identification and Intelligence System) and SOUNDEX versions of the street name, employing two phonetic encoding packages used to eliminate the effects of common spelling errors (Statistics Canada 1989d); a duplicate street name flag (on the AMF) to identify situations where a street name was not unique; a unidirectional street flag (on the AMF) to identify streets that had only a single street direction coded; and an official street name flag (on the AR) to indicate that the street name matched an official AMF street name. The AMF records contained only street data so we appended the Census Subdivision name and a province code and then attempted to assign postal codes to blockface civic numbers. When the postal codes differed between the "from" and "to" civic numbers, we generated subblockfaces for each unique postal code.

For this application, three distinct pockets were created for each record, effectively triplicating the files. The primary pocket was the most stringent in nature and was designed to find all the good match possibilities quickly in the first pass of the files. It was composed of street name/Forward Sortation Area (FSA)/odd or even civic number flag. The second pocket was postal code/odd or even civic number flag which allowed for poorly parsed addresses to be matched on postal code. The third was the NYSIIS version of the street name/odd or even civic number flag which allowed records with spelling variations in street name and missing postal codes to be considered as potential matches.

The function rules established for partial matches for street name, municipality name and the last three characters of the postal code were taken directly from our existing CANLINK application used for internal unduplication where they had already demonstrated their effectiveness.

However, there were three AMFs to which we had difficulty matching in the course of production: Red Deer, St. Thomas and Charny. The problem with all three was missing civic number data on the AMF. Knowing that these would require heavy clerical intervention, a field operation was mounted in December 1990 to update the maps from the Computer Assisted Mapping System (CAM). CAM maps from Geography Division were sent to Regional Office staff who added the missing civic number ranges. These updated maps were subsequently forwarded to Geography Division for inclusion in the next round of updates to the AMF. For the creation of the AR, the civic number ranges for the three AMFs were used manually in the clerical operation.

Success in matching was quite similar across all provinces except for Québec. In Québec, the automatic matching to the blockface dipped by about 10-12% to 73% as it was not as effective at dealing with French addressing as it was with English addressing. Three situations were identified as causes for the drop in the automatic match rate: the use/non-use of articles within the street name (*e.g.*, Savane, de la Savane, la Savane), the use of complete personal names as street names with a high degree of spelling variability (*e.g.*, Jean-François Belanger, J.F. Belanger and Jean F. Belanger) and the lack of street designators. As a result, the clerical operations described below, especially the first one, were of increased importance for matching in Québec relative to the other provinces.

During the AR/AMF processing with the CANLINK software, the only problem that arose was in exceeding an internal pocket maximum on the number of records allowed. The solution was to identify the streets causing the problem from the pocket report (they were always major thoroughfares) and set up special pre-processing programs that would add the fifth digit of the postal code in calculating the pocket value for those streets to make it more discriminating. This had the effect of reducing the number of records within the pocket.

4.6 AR/PCCF Linkage (Step 8)

This step (step 8) attempted to obtain an automated link to the proper Enumeration Area (EA) for those addresses which could not be matched to the blockface using the AMF in step 7.

The principal inputs were the Postal Code Conversion File (PCCF), which gave the correspondence between postal codes and 1986 EAs, and the 1986 to 1991 EA Correspondence File. By matching the two together we could identify postal codes that were uniquely matched to a single 1991 EA, as well as postal codes matched to two or more possible 1991 EAs, requiring manual work to resolve later in step 12.

Again, Brampton became the test vehicle. The analysis of the postal code/EA matching revealed that 38% of the postal codes could be uniquely assigned to a 1991 EA. The linkage to these postal codes of the AR records unmatched to a blockface yielded a further 5% increase in total matches. Overall, the automated match rate increased to 89% (84% to the blockface and 5% to the EA), up from 64% in the September 1989 test, almost cutting in half the amount of manual intervention.

4.7 Loading the Base (Step 9)

To facilitate queries and in anticipation of future usage, ORACLE had been used in the 1989 test as the database management system and was used again for the 1991 production. The ORACLE load step (step 9) involved the transformation of the up-to-now sequential file into four separate component files, one for each of municipality, blockface, street and address.

4.8 Clerical Procedures (Steps 10, 11 and 12)

The clerical procedure for the 1989 test was a review of all unique combinations of street name/street designator/street direction from both AMF and AR records along with an AR record count for each street combination. The objective was to replace an unmatched AR street combination with the legitimate AMF combination. By comparing similar street combinations and determining which ones should in fact have been identical, hitherto uncoded AR records could be matched manually to a particular blockface. This procedure had worked well in 1989 and had proved useful in two problem situations: those where there were large discrepancies in street name spelling and those where the AR street name field contained both the street name and a street designator short form that the PAAS software had not understood in parsing the address.

We expanded the capability of this clerical procedure (step 10) to compare AR street combinations with other similar AR street combinations to handle instances where a particular street might have a number of AR spelling variations with no AMF equivalent. This expansion permitted some additional manual coding of addresses to blockface.

To summarize, in this first clerical procedure (Clerical-1), all addresses not coded automatically to blockface in step 7 (that is, those coded automatically to EA in step 8 and those not yet coded) were examined for possible manual coding to blockface.

Following the Clerical-1 procedure, we added a Compress step (step 11), which was applied to all records coded to the blockface. For each unique value of street name/street designator/street direction within a worksite, all the corresponding address records were checked for uniqueness using the civic number/apartment number as the key. Where multiple records occurred, they were collapsed with all pertinent data blended into one single record, a further step of unduplication.

As a result, at the end of step 10, the database contained addresses coded automatically or manually to blockface, automatically to EA or uncoded as yet.

Step 12 now dealt with those residual addresses that could not be linked to a unique EA but could be matched to two or more possible EAs via step 8. A complete set of CAM-generated maps was produced for the AR project. The Clerical-2 step consisted of examining these maps for the candidate EAs to assign these residual addresses to the proper EA wherever possible.

Overall, the ratio of automated to manual matching was 91%-9%. The automated portion comprised 87% from the AR/AMF linkage to blockface, and 4% from the AR/PCCF linkage to EA. The manual portion was split 3% matched to the blockface from the Clerical-1 operation and 6% to the EA in Clerical-2.

Although ORACLE was an appropriate vehicle for the 1989 prototype, it proved to be costly and eventually a bottleneck once in full production with the AR as just one user on a Bureau-wide database. It allowed for only 8-10% of the worksites on-line at any one time, and had to export and import sites continuously to free up space and reload to carry on processing. A second ORACLE database was therefore set up for exclusive use of the AR team. In fairness to ORACLE, not all the processing being done was conducive to any database management system. The product was being built and as a consequence large portions of the tables were being examined to make sweeping field changes, to eliminate duplication and to select records for printing. ORACLE did offer tremendous flexibility to change software procedures quickly and generate new ones as production unfolded.

4.9 Use of the Computer Assisted Mapping System (Step 13)

The Computer Assisted Mapping System (CAM) was a new research initiative for the 1991 Census whose development ran concurrently with AR development. The system generated all the Enumeration Area maps within AMF coverage areas. This was a major departure from the manual map generation process of the past. CAM also provided a structure to EAs that located blockfaces within blocks and sequenced the blocks within the EA (step 13). An off-shoot to CAM for AR purposes was set up to sequence the dwellings on the blockface. This was necessary to organize the address lists in a manner corresponding more closely to the way the Census Representatives do their listing.

CAM was fully implemented by the time of AR production. In order to remain compatible with it, the same vintage of the AMF that CAM employed was used. However, a small portion of blockfaces had no structure data assigned to them. For any EA where this percentage was greater than 5%, either CAM was re-executed for that worksite if time permitted or an alternate system, Point-in-Polygon Assignments (PIPA), that locates blockfaces within their EA was executed. Although PIPA shifted addresses from the structured portion of the AR booklet (based on blockface coding) to the unstructured portion (EA coding), at least the affected addresses were not dropped during the print selection process, which was the case when sequencing data were missing.

4.10 Printing and Booklet Production (Step 14)

The last production step was the printing and gathering of booklets (step 14) for the almost 23,000 Enumeration Areas containing at this point 6.6 million addresses. Major concerns which were addressed included print speed and quality (a continuous-page printer was used), durability of booklets (the booklets had front and back covers and were stapled) and compilation costs (the booklets were gathered and attached in-house).

5. POST-CENSAL EVALUATION

The post-censal evaluation can be broadly categorized into four study areas: field operations, data capture of AR booklets, update of the AR and determination of the AR contribution to coverage improvements.

Evaluation of field operations will focus on the effectiveness of training, how complete the reconciliation work was, and causes of errors, with a view to improving the methodology for future Censuses.

The data capture operation will yield two separate outputs. First, addresses printed in the booklets will be deleted if invalid, and if valid their Census Household Number will be captured. Second, the new addresses added by the Census Representatives will be captured. It will then be possible to calculate the AR overcoverage and undercoverage rates and the AR contribution to Census coverage. Addresses placed in the wrong EA can be investigated and traced back to the source of error. Through the Census Household Number, the number of persons added and characteristics of dwellings and persons can be studied.

From a cost perspective, the unit cost per dwelling added by the AR will be calculated, in view of the cost of creating the AR and using it in the Census.

6. FUTURE DIRECTIONS

The Address Register (AR), although initially set up as one of the procedures for reducing Census undercoverage, is a developmental project with potential impact on other programs within Statistics Canada as well as other government agencies.

The more immediate objectives for the future development of the AR are as follows: to incorporate the addresses identified during Census enumeration; to evaluate the effectiveness of the AR in improving coverage of the 1991 Census; to document and evaluate the production activities; and to develop a longer-term plan for the AR addressing its cost-effectiveness as a household frame, the optimal updating strategy and its potential for use by external agencies.

Within these guidelines, a project plan was prepared and is presented below under six main topic areas.

6.1 Relationships between the Census and the Address Register

Besides the potential for coverage improvement, other ways in which the AR could contribute to the Census will be explored. Some preliminary thoughts in this regard include possibilities for the AR to be used as a processing control file, for telephone numbers to be used for follow-up purposes, for creation of control numbers of dwellings in an Enumeration Area, for certification of dwelling counts for processing, or for migration analysis. Consideration will be given to whether the AR should be used before or after Census Day, and to how the AR might be used for those addresses where only a higher level of geography than the EA can be ascertained.

6.2 Relationships between Geography and the Address Register

As is evident in the description of the methodology, the creation of the AR relied heavily on many of the products from Geography Division (*e.g.*, the Area Master File, the Postal Code Conversion File). Their contributions and limitations in building the AR will be reviewed. For any new products developed by Geography Division, their possible use in the AR will be investigated with a view to incorporating the AR needs directly into the new product. As well, the AR will be integrated into the Geography Division's Geographical Information System (GIS).

The AR may be able to provide update indicators to the Area Master File (AMF) or for the delineation of Enumeration Areas. The AR could be used to establish priorities especially in high-growth areas or in areas where there are poor civic number ranges in the AMF. The updating of the Postal Code Conversion File might be served by postal code/Enumeration Area or postal code/blockface combinations from the AR. After each Census, all Census households are encoded with blockface centroids. Since the bulk of AR records have already been geocoded prior to the Census, a link of the AR with the Census Household Number will reduce the amount of manual geocoding work after the Census. This last project is already in progress.

6.3 Documentation, Evaluation and Improvement of Procedures

A user guide documenting procedures and a technical guide to document programs, sample problems and solutions and quality assurance are being prepared for the work done to date.

As with any new project, much is learned during the creative process and procedures are developed as required and as time and budget permit. After the fact, there are usually efficiencies to be gained by reviewing these procedures.

For the automated procedures, projects already underway include a more efficient use of ORACLE or choice of another system, the use of desk-top computers rather than the Statistics Canada mainframe computer, standardization of the filter, enhancements to PAAS, amalgamation of sites into provincial databases, the dropping of some fields earlier in the process, consideration of other postal coding software, improvement of address place name matching and an improvement of the Area Master File linkage with French addresses.

For the manual procedures, improved handling of adjacent Enumeration Areas across boundaries of Federal Electoral Districts and of the lack of civic numbers on CAM maps are to be pursued. The editing system to correct addresses will be reviewed for possible improvement as well.

Telephone numbers were added at a later stage within the AR production. A thorough evaluation of their coverage and accuracy will be undertaken especially in view of the potential uses of telephone numbers in the Census and other Statistics Canada surveys. For the latter, initial emphasis will be placed on testing within the context of the upcoming redesign of the Labour Force Survey.

Computer systems developed for the initial production have already been cleaned up to a large extent for better efficiency of mainframe expenditures, for programs and disk and tape storage, for file manipulation, for output, libraries and file access. Better system controls will be prepared.

This AR was produced only for urban areas. Future methodological development will examine the potential for extension to rural areas.

6.4 Updating Methodology

The AR was created from among four sets of administrative files: telephone files, municipal assessment files, hydro files and the T1 tax file from Revenue Canada. As well, the AR is currently being updated to be consistent with the 1991 Census so that the Census is also a source. The relative contributions of these source files, both in volume and quality, will be investigated so that a decision on acquisition of files for updating can be made.

An integral part of the updating strategy is the development of a methodology for updating. The definition of an update will be needed along with an update system. The cost effectiveness of ongoing updating, dependent on the various needs which result from projects identified throughout these future directions, will be considered as well. Is ongoing updating cost effective

when compared to updating only in time for the Census? What requirements will there be from other possible uses? Answers to these questions will lead to an updating strategy.

6.5 Other Uses of the Address Register in Statistics Canada

Besides the Census and geographical relationships presented earlier, a number of other uses are suggested within Statistics Canada. The potential use of the AR in the Labour Force Survey (LFS) will be investigated as part of the LFS Redesign Project. The possibility of using the AR in urban areas either to improve sampling under the existing area frame or as a list frame to reduce the number of stages in the sample design are two major areas highlighted for research. With telephone numbers on the AR, more telephone interviewing would be possible.

The use of the AR as a survey frame for other Statistics Canada surveys will be examined. In addition, since the AR currently uses telephone files as a primary source of information, it has these files on hand for further exploitation. The Special Surveys Program, the General Social Survey and the existing Labour Force Survey are areas which use or require telephone files.

Another potential application within Statistics Canada is as a housing database if the AR were enriched with housing data from the 1991 Census and data obtained from municipal assessment files, for example. The existence of such a database might reduce the amount of information on housing that would have to be collected in future Censuses. Data needs and availability have to be explored.

6.6 Uses of the Address Register External to Statistics Canada

If the AR is to be used outside Statistics Canada, issues of confidentiality of the source files and releasability of the AR must be addressed and meet the requirements of the Statistics Act. Some source files were provided to Statistics Canada in confidence, either contractually (*e.g.*, some files from Alberta) or legally (the T1 file from Revenue Canada).

6.7 Conclusion

The breadth and diversity of the ideas contained above in future directions demonstrate the potential of the Address Register as a geographical product with applications in many areas of Statistics Canada and elsewhere.

ACKNOWLEDGEMENTS

The authors would like to thank the many persons from the following areas for their dedication and perseverance in the creation of the Address Register: Phillip Reed and the AR Production Unit, Geography Division, the Labour Force Survey Sample Control Unit, Census Methodology, Survey Operations Division, the Main Computer Centre and Household Surveys Division. The authors would also like to thank the referee, Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed and Carol Sol for their helpful suggestions for this paper.

REFERENCES

- BOOTH, J.K. (1976). A summary report of all address register studies completed to date. Report E-414E, Statistics Canada.

- DICK, P. (1990). Address register – September 1989 test. Draft internal report, Statistics Canada.
- DREW, J.D., ARMSTRONG, J.B., and DIBBS, R. (1987). Research into a register of residential addresses for urban areas of Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 300-305.
- DREW, J.D., ARMSTRONG, J., VAN BAAREN, A., and DEGUIRE, Y. (1988). Methodology for construction of address registers using several administrative sources. *Proceedings of the Symposium on the Statistical Uses of Administrative Data*, Statistics Canada, 181-190.
- FELLEGI, I.P., and KRÓTKI, K.J. (1967). The testing program for the 1971 Census in Canada. *Proceedings of the Social Statistics Section, American Statistical Association*, 29-38.
- FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- GAMACHE-O'LEARY, V., NIEMAN, L., and DIBBS, R. (1987). Cost implications of mail-out of Census questionnaires using an address register. Internal report, Statistics Canada.
- HILL, T., and PRING-MILL, F. (1985). Generalized iterative record linkage system. *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.
- ROYCE, D. (1986). Address register research for the 1991 Census of Canada. *Journal of Official Statistics*, 2, 4, 447-455.
- ROYCE, D. (1987). Applications of an address register in the Canadian Census. *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada, 207-215.
- ROYCE, D., and DREW, J.D. (1988). Address register research: Current status and future plans. Internal report, Statistics Canada.
- STATISTICS CANADA (1988). Area Master File (AMF), User guide. Statistics Canada.
- STATISTICS CANADA (1989a). Automated Postal Coding System (PCODE), User and retrieval guide. Statistics Canada.
- STATISTICS CANADA (1989b). Generalized Iterative Record Linkage System, Concepts guide. Statistics Canada.
- STATISTICS CANADA (1989c). Postal Address Analysis System (PAAS), User guide. Statistics Canada.
- STATISTICS CANADA (1989d). Record linkage software, Reference guide. Statistics Canada.
- STATISTICS CANADA (1990). *User's guide to the quality of 1986 Census data: Coverage*. Catalogue 99-135E, Statistics Canada.
- STATISTICS CANADA (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide. Statistics Canada.
- SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register at Statistics Canada. *Proceedings of the Symposium on Spatial Issues in Statistics*, Statistics Canada.
- VAN BAAREN, A. (1988). Report on the November 1987 address register test. Internal report, Statistics Canada.

Bibliography on Capture-Recapture Modelling With Application to Census Undercount Adjustment

STEPHEN E. FIENBERG¹

ABSTRACT

This article presents a selected annotated bibliography of the literature on capture-recapture (dual system) estimation of population size, on extensions to the basic methodology, and the application of these techniques in the context of census undercount estimation.

KEY WORDS: Capture-recapture; Census undercount; Dual system estimation; Loglinear models.

1. INTRODUCTION

The method of capture-recapture for estimating the size of a closed population has been in use since at least the nineteenth century, when Peterson (1896) developed the standard estimator that bears his name for the use with fish populations. Subsequent application to other types of populations include Geiger and Werner (1924) – physics; Lincoln (1930) – wildlife; Chandrasekar and Deming (1948) – vital statistics for human populations; Wittes and Sidel (1968), Wittes, Colton and Sidel (1974) – epidemiology; Sanathanan (1972b) – particle scanning in physics; Blumenthal and Marcus (1975) – life testing; Green and Stollmack (1981), Rossmo and Routledge (1990) – crimes and criminals. In the context of the study of human populations and demography the method is often referred to as dual system estimation. We have included virtually no references to the related problem of counting the number of species, which goes back to the work of R.A. Fisher in the 1940s and had an elegant formulation in Efron and Tibshirani's (1976) *Biometrika* paper on "How many words did Shakespeare know?".

The basic capture-recapture approach rests on a number of assumptions, *e.g.*: (1) the population under study is closed; (2) individuals (units) can be perfectly matched from capture to recapture; (3) capture probabilities are constant across the individuals (units) in the population; (4) the probability of inclusion of an individual (unit) in recapture sample is independent of inclusion in original census or sample. Beginning in the late 1930s various investigators began to explore extensions that allowed for departures from the assumptions. These methods typically require additional data such as a second recapture (or even a third) and the full capture-recapture history of each individual.

For human populations and the study of vital statistics the methodology has long been linked to census data, *e.g.*, see Tracy (1941) and Shapiro (1949, 1954). In connection with the 1950 decennial census of population, the U.S. Bureau of the Census introduced the use of a sample matched to the census records for coverage evaluation. This approach has evolved into what is currently known as the Post Enumeration Survey approach to undercount and overcount estimation, and it has been the focal point of the recent and ongoing controversy of the possible adjustment of the 1980 and 1990 censuses, *e.g.*, see Eriksen and Kadane (1985); Freedman and Navidi (1986, 1992); Freedman (1991); Wolter (1991).

¹ Stephen E. Fienberg, Office of Vice President (Academic Affairs), York University, North York, Ontario M3J 1P3, Canada.

This selected annotated bibliography presents an overview of published literature on capture-recapture estimation of population totals. It includes historical references, articles that explore departures from assumptions and extensions of the basic methodology, and is most complete in connection with papers that describe the dual and multiple system approaches in the context of census undercount estimation. In this regard, however, we have not included references to any of the unpublished memoranda and papers from the U.S. Bureau of the Census (primarily because most of these have been replicated in some form in the published literature). We have tended to exclude articles published in unrefereed proceedings for related reasons. Because the literature on specialized applications of capture-recapture techniques to wildlife populations is so extensive, and only some of it is of relevance for human populations, we have provided primarily references to reviews of this literature, *e.g.*, see Brownie *et al.* (1977); Otis *et al.* (1978); Seber (1973, 1982). Similarly we have included only a small number of references to the more specialized methods in use for life testing, *e.g.*, see Dahiya and Blumenthal (1986), as well as those in use for software reliability applications, *e.g.* Jelinski and Moranda (1972), and Duran and Wiorkowski (1981). The methods in this latter literature diverge in significant ways from those used in the basic capture-recapture and dual system approaches.

2. SELECTED BIBLIOGRAPHY

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

- Extends usual dual systems approach to allow for multiplicative stratification effects.

BAKER, S. G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates (with discussion). *Biometrics*, 46, 1193-1200.

- Links cross-classification of covariates to the capture and recapture via loglinear models and then uses EM algorithm to estimate population size.

BIEMER, P.P. (1988). Modelling matching error and its effect on estimates of census coverage error. *Survey Methodology*, 14, 117-134.

- Develops models for evaluating impact of matching error on census coverage.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.H. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Chapter 6. Cambridge, MA: MIT Press.

- Monograph on loglinear models which includes a chapter on the relationship to capture-recapture models.

BLUMENTHAL, S., and MARCUS, R. (1975). Estimating population size with exponential failure. *Journal of the American Statistical Association*, 70, 913-922.

- Uses exponential distribution to estimate population size based on a subset of observations obtained by truncated sampling.

BOSWELL, M.T., BURNHAM, K.P., and PATIL, G. P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies. In *Handbook of Statistics 6: Sampling*, (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: North Holland, 469-488.

- Gives succinct summary of several basic variants on capture-recapture models and their estimation.

BROWNIE, C., ANDERSON, D.R., BURNHAM, K.P., and ROBSON, D.S. (1977). Statistical inference from band recovery data: a handbook. *U.S. Fisheries and Wildlife Service Resource Publication No. 131*.

- Describes a comprehensive range of capture-recapture models and appropriate goodness-of-fit tests, with emphasis on banding experiments.

BURGESS, R.D. (1988). Evaluation of reverse record check estimates of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.

- Describes the survey-based accounting approach of the reverse record check for undercount estimation. Does not deal with issue of exclusion of individuals from census and other lists.

BURNHAM, K. P., ANDERSON, D.R., WHITE, G.C., BROWNIE, C., and POLLOCK, K.H. (1987). *Design and Analysis Methods for Fish Survival Experiments Based on Release-Recapture*. Bethesda, MD: American Fisheries Society.

- Combines methodology of Brownie *et al.* for band recovery with survival estimation under Jolly-Seber mark-recapture models.

BURNHAM, K.P., and OVERTON, W.S. (1978). Estimation of the size of a closed population when the capture probabilities vary among animals. *Biometrika*, 65, 625-633. Correction (1981) 68, 345.

- Develops a capture-recapture model with heterogeneity for animals but constant probabilities of capture across samples. Model induces dependencies amongst captures.

CASTELDINE, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

- Develops a Bayesian approach using beta priors for traditional independence-based Schnabel census model for multiple recapture data.

CHAKRABORTY, P.N. (1963). On a method of estimating birth and death rates from several agencies. *Calcutta Statistical Association, Bulletin*, 12, 106-112.

- Extends Chandrasekar-Deming approach to three or more sources.

CHANDRASEKAR, C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

- Develops dual-system technique and suggests the use of stratification for eliminating heterogeneity. Applies approach to estimation of number of births and deaths in several Indian villages.

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catch ability. *Biometrics*, 43, 783-791.

- Explores heterogeneous catchability model of Burnham and Overton using a moment inequality to get a lower bound on population size.

CHAO, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45, 427-438.

- Explores adequacy of estimator resulting from moment inequality for heterogeneous catchability model in settings involving sparse data.

CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.

- Develops the hypergeometric sampling model for estimating the population size in capture-recapture studies.

CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian census count for under enumeration. *Survey Methodology*, 14, 173-189.

- Describes the use of dual system estimation and a post enumeration survey to adjust the results of the Australian census. Also applies Wolter sex-ratio technique to check on sensitivity of dual system estimator.

CHRISTENSEN, H.T. (1958). The method of record linkage applied to family data. *Marriage and Family Living*, 20, 38-43.

CITRO, C. F., and COHEN, M. L., (Eds.) (1985). *The Bicentennial Census. New Directions for Methodology* in 1990. Washington, DC: National Academy Press.

- Report of a panel of the Committee on National Statistics on census methodology including an examination of the dual systems approach to undercount correction.

COALE, A.J. (1961). The design of an experimental procedure for obtaining accurate vital statistics. *International Population Conference*, New York, 372-375.

- Proposes the use of two lists covering the same sample from a population.

COHEN, M.L. (1990). Adjustment and reapportionment – analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.

- Examines effect of bias and variability on accuracy of adjusted and unadjusted census counts and the impact on the reapportionment of the U.S. House of Representatives.

CORMACK, R. M. (1981). Loglinear models for capture-recapture experiments on open populations. In *The Mathematical Theory of the Dynamics of Biological Populations*, II (Eds. R.W. Hiorns and D. Cooke). London: Academic Press, 217-235.

- Introduces Poisson model for capture-recapture and uses it with loglinear models to extend standard approach to allow for birth, death, and trap dependency.

CORMACK, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.

- Uses Poisson model and loglinear representation for inclusion of birth, death, and trap dependency into standard capture-recapture approach.

CORMACK, R. M., and JUPP, P.E. (1991). Inference for Poisson and multinomial models for capture recapture experiments. *Biometrika*, 78, 911-916.

- Compares MLEs of parameters under the two models and presents relationship between the corresponding asymptotic variances and covariances.

COWAN, C.D., and MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

- Extends dual systems approach to situation involving clustered observations as in the U.S. census coverage improvement program.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 191-208.

- Proposes a PES-based model for undercount adjustment utilizing an empirical Bayes estimation scheme and a family of loss functions.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

- Develops and applies empirical Bayes smoothing methods for census adjustment factors produced from dual systems approach for geographic by demographic stratification. Applies approach to state data from 1980 U.S. census.

CRESSIE, N., and DAJANI, A. (1991). Empirical Bayes estimation of U.S. undercount based on artificial populations. *Journal of Official Statistics*, 7, 57-67.

- Shows that synthetic estimation approach used by Isaki *et al.* is special case of empirical Bayes.

CROXFORD, A.A. (1968). Record linkage in education. In *Record Linkage in Medicine* (Ed. E.D. Acheson). London: E. and S. Livingstone, 351-356.

- DAHIYA, R.C., and BLUMENTHAL, S. (1986). Population or sample size estimation. In *Encyclopedia of Statistical Sciences*, (Volume 7), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 100-110.
- Reviews theory underlying population size estimation from truncated sampling for discrete distributions and provides references to domains of application.
- DARROCH, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.
- Describes the maximum likelihood approach to the multiple recapture problem under complete independence.
- DARROCH, J.N. (1959). The multiple-recapture census II: Estimation when there is immigration or death. *Biometrika*, 46, 336-351.
- Extends the maximum likelihood approach under independence to open populations with either immigration or death.
- DARROCH, J.N. (1961). The two sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, 45, 343-359.
- Extends the maximum likelihood approach with independence to the situation where the original captured individuals are stratified into s groups and the individuals in the recapture sample are stratified, but according to t (possibly different) strata.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1992). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. Submitted for publication.
- Extends triple system estimation to allow for individual heterogeneity and selected forms of dependence. Applies estimators to triple system data from census dress rehearsal in St. Louis.
- DARROCH, J.N., and RATCLIFF, D. (1980). A note on capture-recapture estimation. *Biometrics*, 36, 149-153.
- Presents an alternative estimator for capture-recapture problems with interesting asymptotic properties.
- DASGUPTA, P. (1964). On the estimation of the total number of events and of the probabilities of detecting an event from information supplied by several agencies. *Calcutta Statistical Association, Bulletin*, 13, 89-100.
- Extends Chandrasekar-Deming approach to three or more sources.
- DAVIDSON, L. (1962). Retrieval of misspelled names in an airline passenger record system. *Communications of the Association of Computer Machinery*, 5, 169-171.
- DEMING, W.E., and KEYFITZ, N. (1967). Theory of surveys to estimate total populations. In *Proceedings of the World Population Conference*, Belgrade, 1965 (Vol. 3). New York: United Nations, 141-144.
- Extends of Chandrasekar-Deming approach to three sources.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology*, 14, 71-86.
- Describes implementation of post-enumeration survey approach to dual system estimation in a test census.
- DING, Y. (1990). Capture-Recapture Census with Uncertain Matching. Ph.D. dissertation, Department of Statistics, Carnegie Mellon University.
- Develops a probabilistic matching model for use with dual and multiple system estimation, and considers a Bayesian approach for estimating the population size. Illustrates techniques using data from test census results from Los Angeles.

DING, Y., and FIENBERG, S.E. (1992). Estimating population and census undercount in the presence of matching error. Submitted for publication.

- Develops a probabilistic matching model for use with dual system estimation and illustrates its application to data from test census results from Los Angeles.

DURAN, J.W., and WIORKOWSKI, J.J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Systems Engineering*, 7, 147-148.

EFRON, B., and THISTED, R.A. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63, 435-467.

- Adapts a parametric model due to Fisher and a nonparametric model for the classical species problem using empirical Bayes methods. Applies approach to the vocabulary of Shakespeare.

EL-KHORAZATY, M.N., IMREY, P.B., KOCH, G.G., and WELLS, H.B. (1977). Estimating the total number of events with data from multiple record systems: a review of methodological strategies. *International Statistical Review*, 45, 129-157.

- Review of literature and methods for dual- and multiple systems estimation. Includes sections comparing use of techniques and departures from assumptions in wildlife and human populations.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

- Applies dual system approach to 1980 census data, including the regression-based smoothing of undercount estimates and the estimation of adjusted odds ratios using demographic estimates.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.

- Presents revisions and extensions to the Ericksen and Kadane methodology and a critique of Freedman and Navidi.

FAY, R. E., PASSEL, J. S., ROBINSON, J. G., and COWAN, C. D. (1988). *The Coverage of Population in the 1980 Census*. Bureau of the Census. Washington, DC: U. S. Department of Commerce.

- Official Bureau of the Census report on attempts to measure undercount in the 1980 U.S. decennial census.

FEIN, D.J., and WEST, K.K. (1988). The sources of census undercount: Findings from the 1986 Los Angeles Test Census. *Survey Methodology*, 14, 223-240.

- Attempts to test hypotheses regarding the causes of census undercount for a hard-to-enumerate Hispanic urban population.

FIENBERG, S.E. (1972). The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika*, 59, 591-603.

- Introduces a method for estimating dependencies among multiple lists using loglinear models and develops a general approach for estimation using results on 2^k incomplete contingency tables and conditional estimation.

FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. In *Encyclopedia of Statistical Sciences*, (Supplemental Volume), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 181-185.

- Presents historical background on the differential undercount of the U.S. population and brief descriptions of demographic analysis and the dual system estimation approaches.

FREEDMAN, D. A. (1991). Policy forum: Adjusting the 1990 census. *Science*, 252, 1233-1236.

- Critique of dual systems approach to adjustment of the 1990 census.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models and adjusting the 1980 census (with discussion). *Statistical Science*, 1, 3-39.

- Critique of Ericksen and Kadane dual systems methodology as applied to 1980 census data.

FREEDMAN, D.A., and NAVIDI, W.C. (1992). Should we have adjusted the census of 1980? (with discussion). *Survey Methodology*, this issue.

- Continues critique of the use of dual system estimation and synthetic adjustment as applied to 1980 census.

GARTHWAITE, P.H., and BUCKLAND, S.T. (1990). Analysis of multiple-recapture census by computing conditional probabilities. *Biometrics*, 46, 231-238.

- Uses a recursive relationship to generate point and interval estimate for multiple-recapture census under independence.

GEIGER, H., and WERNER, A. (1924). Die Zahl der ion radium ausgesandten α -Teilchen. *Zeitschrift für Physik*, 21, 187-203.

- Applies a capture-recapture method to radium ion particle detection estimation.

GOLDBERG, J.D., and WITTES, J.T. (1978). The estimation of false negatives in medical screening. *Biometrics*, 34, 77-86.

- Applies capture-recapture models to problems in medical screening.

GOUDIE, I. B. J. (1990). A likelihood-based stopping rule for recapture debugging software reliability. *Biometrika*, 77, 203-206.

GREEN, M.A., and STOLLMACK, S. (1981). Estimating the number of criminals. In *Models in Quantitative Criminology*, (Ed. J.A. Fox). New York: Academic Press, 1-24.

GREENFIELD, C.C. (1975). On the estimation of a missing cell in a 2×2 contingency table. *Journal of the Royal Statistical Society, Series A*, 138, 51-61.

- Introduces a non-zero value for the response correlation, by taking the mid-point of the range of permissible correlation values, and consequently derives a value for missing cell. Applies approach to census data from Malawi.

GREENFIELD, C.C. (1976). A revised procedure for dual record systems in estimating vital events. *Journal of the Royal Statistical Society, Series A*, 139, 389-401

- Applies bounds on correlation in a 2×2 table to dual system estimation in the presence of event correlation induced by heterogeneity.

GREENFIELD, C.C., and TAM, S.M. (1976). A simple approximation for the upper limit to the value of a missing cell in a 2×2 contingency table. *Journal of the Royal Statistical Society, Series A*, 139, 96-103.

- Uses approximation for upper bound for response correlation to derive an upper bound for missing cell.

HOGAN, H., and WOLTER, K.M. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

- Reports on Los Angeles Test of Adjustment Related Operations and estimates of sources of bias in post enumeration survey and census-based dual systems estimates.

HOLST, L. (1973). Some limit theorems with applications in sampling theory. *Annals of Statistics*, 1, 644-658.

- Applies results on successive sampling to derive asymptotic distribution of usual Peterson estimator when there are heterogeneous capture probabilities or the effects of matching.

HOOK, E., and REGAL R. (1982). Validity of Bernoulli census, log-linear, and truncated binomial models for correcting underestimates in prevalence studies. *American Journal of Epidemiology*, 116, 168-176.

- Applies different loglinear related methods used to study the number of infants born with Downs syndrome.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

- Uses linear logistic models for capture probabilities for individuals and capture occasions.

HUGGINS, R.M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47, 725-732.

- Uses linear logistic models for capture probabilities and exploits temporal order of captures to introduce dependence amongst captures and on measurable covariates for those captured at least once.

ISAKI, C.T. (1986). Bias of the dual system estimator and some alternatives. *Communications in Statistics, Theory and Methods*, 15, 1435-1450.

- Exploits upper bound on correlation bias to reduce the bias of the dual system estimator.

ISAKI, C.T., and SCHULTZ, L.K. (1986). Dual system estimation using demographic analysis data. *Journal of Official Statistics*, 2, 169-179.

- Uses demographic analysis data to get revised dual system estimates for 1980 census using different models for correlation bias.

ISAKI, C.T., and SCHULTZ, L.K. (1987). The effect of correlation and matching error in dual system estimation. *Communications in Statistics, Theory and Methods*, 16, 2405-2427.

- Develops a simple matching error model in the presence of correlation bias to compare three dual system estimators.

ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., and HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.

- Develops simulation populations based on 1980 census and coverage evaluation results, evaluates regression-based synthetic undercount estimation methods, and shows superiority of synthetic approaches to raw census counts.

JABINE, T.B., and BERSHAD, M.A. (1968). Some comments on the Chandrasekar and Deming technique for the measurement of population change. Paper presented at CENTO Symposium on Demographic Statistics, Karachi, Pakistan.

- Shows that positive correlation bias produces a downward bias in estimate of total population size.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Test Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

- Describes census methodology for matching census and post-enumeration survey records, with the results from their application to 1985 test census.

JELINSKI, Z., and MORANDA, P.B. (1972). Software reliability research. In *Statistical Computer Performance Evaluation*, (Ed. W. Freiburger). New York: Academic Press, 465-484.

- Proposes a model with exponentially distributed failures to estimate total number of program faults based on times of occurrence of failures in fixed time period.

JEWELL, W.S. (1985). Bayesian estimation of undetected errors. In *Bayesian Statistics 2*, (Eds. J.M. Bernardo, *et al.*). New York: Elsevier, 663-671.

JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration – stochastic models. *Biometrika*, 52, 225-247.

- Estimation from multiple-recapture data for open populations.

KADANE, J.B., MEYER, M.M., and TUKEY, J.W. (1992). Correlation bias in the presence of stratum heterogeneity. Submitted for publication.

- Demonstrates the impact of correlation bias resulting from collapsing over heterogeneous strata with different catchability probabilities in each strata, subject to a monotonicity constraint.

KRÓTKI, K.J. (Ed.) (1978). *Developments in Dual System Estimation of Population Size and Growth*. Edmonton: University of Alberta Press.

- Reviews the use of dual system estimation for vital records in various countries. Includes technical details on the use of complex samples and elaborations on basic techniques.

LASKA, E.M., MEISNER, M., and SIEGEL, C. (1988). Estimating the size of a population from a single sample. *Biometrics*, 44, 461-472. Correction, (1989), 45, 1347.

- Estimates population size from the last of k lists.

LEWIS, C.E., and HASSANEIN, K.M. (1969). The relative effectiveness of different approaches to the surveillance of infection among hospitalized patients. *Medical Care*, 8, 379-384.

- Applies dual system estimation to surveillance of infectious diseases.

LINCOLN, F.C. (1930). Calculating waterfowl abundance on the basis of banding returns. *Circular of the U.S. Department of Agriculture*, 118, 1-4.

- Applies capture-recapture method to estimating size of waterfowl populations.

MANTEL, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics*, 7, 240-246.

- Shows how heterogeneity induces correlation bias (event correlation) in the estimation of disease prevalence.

MARKS, E.S., SELTZER, W., and KRÓTKI, K.J. (1974). *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: Population Council.

- Comprehensive review of dual-systems estimation, assumptions, background, design, and problems. Contains claim that the basic method has been used for more than three centuries for estimating size of animal populations.

MAXIM, L.D., HARRINGTON, L., and KENNEDY, M. (1981). A capture-recapture approach for estimation of detection probabilities in aerial surveys. *Photogrammetric Engineering and Remote Sensing*, 47, 779-788.

MULRY, M. H., and SPENCER, B.D. (1988). Total error in the dual system estimator: the 1986 census of Central Los Angeles County. *Survey Methodology*, 14, 241-263.

- Develops a total error model for dual systems approach applied to Los Angeles Test of Adjustment Related Operations.

MULRY, M. H., and SPENCER, B.D. (1991). Total error in PES estimates of population (with discussion). *Journal of the American Statistical Association*, 86, 839-863.

- Extends earlier Mulry-Spencer development of total error model for dual systems approach and applies it to 1988 dress rehearsal census in St. Louis and east-central Missouri.

NICHOLS, J.D., and POLLOCK, K.H. (1983). Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture-recapture models. *Paleobiology*, 9, 150-163.

OTIS, D.L. BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monograph*, 62, Washington, DC: Wildlife Society.

- Reviews capture-recapture and related methods for wildlife populations.

- PERKINS, W.M., and JONES, C.D. (1965). Matching for census coverage checks. Paper presented at the Meetings of the American Statistical Association, Philadelphia.
- PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. Report of the Danish Biological Station to the Ministry of Fisheries, 6, 1-48.
- Classic development of method of capture-recapture and its application to the estimation of the size of fish populations.
- POLLACK, E.S. (1965). Use of census matching for study of psychiatric admission rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-115.
- RAJ, D. (1977). On estimating the number of vital events in demographic surveys. *Journal of the American Statistical Association*, 72, 377-381.
- Develops a formula for bias in dual system estimator under a general model of response errors and explores use of double sampling to correct bias.
- ROSSMO, D.K., and ROUTLEDGE, R. (1990). Estimating the size of criminal populations. *Journal of Quantitative Criminology*, 6, 293-314.
- RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation strategies for missing values in post-enumeration surveys. *Survey Methodology*, 14, 209-221.
- Presents a methodology matching for undercount estimation which utilizes an imputation approach rooted in loglinear models in the presence of missing data.
- SANATHANAN, L.P. (1972a). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- Demonstrates asymptotic equivalence of conditional and unconditional estimators for the population size.
- SANATHANAN, L.P. (1972b). Models and estimation methods in visual scanning experiments. *Technometrics*, 14, 813-829.
- Develops latent model to estimate the number of particles in scanning records which allows for differential detectability and induces dependencies amongst detectors.
- SANATHANAN, L.P. (1973). A comparison of some models in visual scanning experiments. *Technometrics*, 15, 67-78.
- Applies traditional capture recapture model and latent models to data from actual visual scanning experiments.
- SANDLAND, R.L., and CORMACK, R.M. (1984). Statistical inference for Poisson and multinomial models for capture recapture experiments. *Biometrika*, 71, 27-33.
- Shows relationship between the asymptotic variances of the population size under general capture-recapture model for the two alternate sampling schemes.
- SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 Test of Adjustment Related Operations. *Survey Methodology*, 14, 87-98.
- Examines effect of missing data on dual system estimate applied to test census data.
- SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.
- Applies synthetic estimation approaches to 1980 census data to evaluate the impact of undercount estimation.

SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.

- Extends the basic capture-recapture approach to multiple recaptures, with information at each recapture on whether individuals were captured previously.

SCOTT, C. (1974). The dual record (PGE) system for vital rate measurement, some suggestions for further development. In *International Population Conference, Liege, 1973*, (Volume 2). Liege: International Union for the Scientific Study of Population.

SEBER, G.A.F. (1965). A note on the multiple-recapture census. *Biometrika*, 52, 249-259.

- Estimation from multiple-recapture data for open populations with time-specific parameters.

SEBER, G.A.F. (1973). *The estimation of Animal Abundance and Related Parameters*. New York: Hafner. Second edition (1982). New York: Macmillan.

- Contains an up-to-date review of capture-recapture techniques and their extensions for animal populations, with emphasis on applications.

SEBER, G.A.F. (1982). Capture-recapture methods. In *Encyclopedia of Statistical Sciences*, (Volume 1), (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 367-374.

- Reviews capture-recapture approach for both closed and open populations and provides guide to wildlife and fisheries applications.

SHAPIRO, S. (1949). Estimating birth registration completeness. *Journal of the American Statistical Association*, 45, 261-264.

- Describes use of 1940 U.S. census data to check on completeness of birth registration using Chandrasekar-Deming estimator.

SHAPIRO, S. (1954). Recent testing of birth registration completeness in the United States. *Population Studies*, 8, 3-21.

- Describes use of U.S. census data from 1940 and 1950 to check on completeness of birth registration using Chandrasekar-Deming estimator.

SIRKEN, M. G. (1978). Dual systems estimators based on multiplicity surveys (with discussion). Chapter 4 in *Developments in Dual System Estimation of Population Size and Growth*, (Ed. K. J. Krótki). Edmonton: University of Alberta Press, 81-91.

- Adapts author's approach of multiplicity surveys for rare events to dual systems problem.

SMITH, P.J. (1988). Bayesian methods for capture-recapture surveys. *Biometrics*, 44, 1177-1189.

- Uses Poisson approximation and Gamma prior for a Bayesian approach to estimation under independence in multiple recapture model.

SMITH, P. J. (1991). Bayesian analyses for multiple capture-recapture model. *Biometrika*, 78, 399-407.

- Develops exact Bayesian posterior distribution for multiple recapture census under independence of recaptures.

SRINIVASAN, S.K., and MUTHIAH, S.A. (1968). Problems of matching births identified from two independent sources. *Journal of Family Welfare*, 14, 13-22.

TRACY, W.R. (1941). Fertility of the population of Canada. Reprinted from *Seventh Census of Canada, 1931*, (Vol. 2), Census Monograph no. 3. Ottawa: Cloutier.

- An early application of dual systems approach to census data.

WINKLER, W.E. (1989). Methods for adjusting for lack of independence in an application of the Felligi-Sunter model of record linkage. *Survey Methodology*, 15, 101-117.

- Examines methods of adjusting linkage rules for matching dual-system records when independence cannot be assumed.

WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 93-97.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly and other problems.

WITTES, J.T., COLTON, T., and SIDEL, V.W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly.

WITTES, J.T., and SIDEL, V.W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases*, 21, 287-301.

- Uses capture-recapture approach to estimate number of hospital patients using methicillin.

WOLTER, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

- Describes basic version of dual systems approach as used for the 1980 census, including the elimination of erroneous enumerations.

WOLTER, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

- Uses sex ratio to get a modified capture-recapture estimator for data with stratification by sex, assuming either a common odds ratio, or independence in one stratum. Applies approach to animal data and describes application to census data.

WOLTER, K.M. (1991). Policy Forum: Accounting for America's uncounted and miscounted. *Science*, 253, 12-15.

- Describes 1990 census adjustment procedures and why they are statistically defensible.

WOLTER, K. M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

- Examines through simulations the improvements of synthetic adjustment over the census counts for small areas.

ZALAVSKY, A.M., and WOLFGANG, G.S. (1990). Triple system modelling of census, post-enumeration survey, and administrative list data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.

- Applies various loglinear and related models of dependence to analyze triple system data from census dress rehearsal in St. Louis.

A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure

LINDA K. ROE, JOHN F. CARLSON and DAVID A. SWANSON¹

ABSTRACT

This paper examines the suitability of a survey-based procedure for estimating populations in small, rural areas. The procedure is a variation of the Housing Unit Method. It employs the use of local experts enlisted to provide information about the demographic characteristics of households randomly selected from residential unit sample frames developed from utility records. The procedure is nonintrusive and less costly than traditional survey data collection efforts. Because the procedure is based on random sampling, confidence intervals can be constructed around the population estimated by the technique. The results of a case study are provided in which the total population is estimated for three unincorporated communities in rural, southern Nevada.

KEY WORDS: Survey-based; Utility records; Confidence intervals; Nevada.

1. INTRODUCTION

In its most recent survey of state and local agencies preparing population and housing estimates, the U.S. Bureau of the Census found that about 89 percent of the agencies surveyed use the Housing Unit Method (HUM) (Byerly 1990). This method was also found to be widely used in an earlier survey (U.S. Bureau of the Census 1978). The method has been found to provide accurate estimates of the total population (Lowe, Pittenger and Walker 1977; Lowe, Weisser and Myers 1984; Smith and Lewis 1980, 1983; Smith and Mandell 1984) as well as a strong conceptual and practical foundation for a municipal estimation system (Martin and Serow 1979; Rives and Serow 1984; Swanson, Baker and Van Patten 1983).

One of the strong features of the HUM is that it can be implemented in a variety of forms, which allows it to be adapted to a range of data environments (Swanson, Baker and Van Patten 1983). This adaptability has been exploited primarily by subnational demographic centers for purposes of revenue sharing and related programs (Martin and Serow 1978; Swanson, Baker and Van Patten 1983). However, as pointed out by Rives (1982), the method has potential uses in other arenas.

As an example, consider the case of environmental impact statements. Concerns over legal and environmental issues have resulted in decisions to locate unpopular facilities in sparsely populated rural areas for which census and other socioeconomic data are usually not available (Freudenburg 1982; Brown, Geertsen and Krannich 1989; Munsell 1988). As a consequence, it has become necessary to develop methods of inquiry, particularly suited for small, rural areas, that fully exploit available data, are less costly and, in many cases, less intrusive, than area, telephone, and mail surveys. We believe that the variation of the HUM that we propose in this paper contributes to this type of methodological development.

¹ Linda K. Roe and John F. Carlson, Science Applications International Corporation, 101 Convention Center Drive, Las Vegas, Nevada 89109; David A. Swanson, Center for Social Research and Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447-0003, USA.

The HUM variation that we describe in this paper combines two methods that are, in themselves, well known. However, they have largely been developed in isolation from each other, as well as from the HUM. These are: (1) random sampling; and (2) "local expert" interviews. As discussed later, these methods, combined with the HUM may lead to a means of obtaining the population size and, eventually, composition data required to meet the information needs of impact assessment projects and other activities affecting small, rural areas.

2. CONSIDERATIONS IN ASSESSING IMPACTS IN SMALL, RURAL AREAS

The location of new plants or industries in rural areas generally requires a work force exceeding that which is available in the local area. Population growth in the communities that are in close proximity to the site can be expected to vary according to the size of the project and the number of employees that will be hired to build, then to operate and maintain the completed facility. Whether rapid increases in the overall number of individuals are expected, or significant changes in the age and sex distribution, the altered population structure will have an effect on the type and amount of public services needed (Summers 1982). Thus, impact assessments require information regarding anticipated increases in school enrollment, housing requirements, health care needs, and other services. Before such projections can be made, however, information on the current population in the impacted area must be determined in order to have a "jump-off" or "launch" population for forecasting purposes (Carlson, Williams and Swanson 1990; Pittenger 1976; U.S. Department of Energy 1988).

The understanding of major factors affecting the distribution of people in isolated rural areas is critical in constructing demographic profiles and projections. These communities are likely to have been affected by periods of both boom growth and decline (Krannich and Greider 1984). Historical patterns of population change, as well as current trends, may differ substantially from averages derived from that of the county as a whole or even other sub-county areas. This presents a special problem because accurate demographic information is usually available only for years in which the Federal Census is conducted. However, census data, including information on households, are not typically available for unincorporated places with small populations. Since cost is usually a major factor, the possibility of conducting special censuses or large sample surveys, particularly on a regular basis, is often precluded, even in small, rural areas. An additional problem associated with such counts is that they require interviewers to contact individual households, which imposes on time and privacy and adds to disruption burdens that may be already high for local residents (Brown, Geertsens and Krannich 1989; Krannich, Berry and Greider 1989; Schleifer 1986).

The estimation of the size of the current population of a small, rural area could, in principle, be accomplished through several techniques. However, data limitations and a desire for accuracy severely curtail the range of candidates and, realistically, point to a single technique: HUM (Smith 1986; Smith and Mandell 1984; Lowe, Weissner and Myers 1984; Swanson, Van Patten and Baker 1983; Smith and Lewis 1983, 1980).

3. THE HOUSING UNIT METHOD

The concept of the HUM relies on the fact that nearly everyone sleeps under some kind of shelter. The U.S. Bureau of the Census, for example, chooses to define two classes of shelters: group quarters; and housing units. All persons are assigned to one shelter class or the other. The HUM holds that these shelters can be identified, counted, and classified as occupied or

vacant. Also, all occupied shelters must have a specific number of occupants. Therefore, the population of any given place must be equal to the sum of the housing units times the occupancy rate times the average number of persons per occupied housing unit (household) plus the number of persons in group quarters. The four elements of the HUM provide an exact demographic identity, with the population of a given place given by

$$P = [(H) * (O) * (PPH)] + GQ,$$

where

- P = total population,
- H = total housing units,
- O = proportion of occupied units,
- PPH = mean number of persons per household,
- GQ = group quarters population.

The key accuracy issue in using the HUM is in the determination of each of the components. Moreover, as Smith (1986, pp. 245-246) observes:

“The Housing Unit Method is a robust, comprehensive, and extremely flexible form of population estimation with a number of characteristics that make it useful for small-area analysis. It is not confined to a single technique or type of data; rather, it can utilize a number of different techniques and data sources, including those that may be applicable in one area but not another.”

As also noted by Smith (1986), there are two major approaches used to generate the “number of households” element of the HUM. One relies on measures of construction activity and the estimation of an occupancy rate; the other uses utility data, such as residential electrical customers. A major advantage of the second approach is that it can directly provide the number of households, which eliminates or substantially reduces a number of potential data inaccuracies, including the need to estimate time lags between when permits are issued and units are completed, completion rates, demolitions, conversions, and occupancy rates. Starsinic and Zitter (1968) as well as Rives and Serow (1984) find that the “utility data” approach to the HUM is advantageous, although they also acknowledge certain limitations.

Another advantage of using utility data is that the same data used to obtain total households can also be used as a complete frame from which samples can be drawn in order to obtain an estimate of the average number of persons per household (PPH). There are three forms that traditional data collection usually take in obtaining this type of sample information: mail, telephone, and personal interview. We propose that in their place “local experts” be used to minimize both cost and disruption burdens.

4. LOCAL EXPERTS

The local expert procedure (also referred to as the key informant procedure) of obtaining information about a community is well-established in the field of cultural anthropology. It is generally acknowledged as a “reliance on a small number of knowledgeable participants, who observe and articulate social relationships for the researcher” (Seidler 1974, p. 816). Further, Poggie (1972) finds that when the questions asked in the field relate to noncontroversial, concrete, and directly observable public phenomena, local experts are a highly reliable and precise source of information.

There are two key issues in using the local expert procedure in conjunction with utility records and the HUM. The first is to identify and recruit people who are truly local experts on the composition of the households presented to them in the sample. The second is to be able to obtain household identifying information that is familiar to the local experts (*e.g.*, a street address and the name of the householder instead of a utility company billing code).

5. CASE STUDY

The data collection activity on which our population estimates rely is part of a program to assess the socioeconomic characteristics of communities located near Yucca Mountain, Nevada, the proposed site of a geologic nuclear waste repository (U.S. Department of Energy 1988). The data will comprise part of the set used in a comprehensive impact analysis of the proposed repository.

Yucca Mountain is located in Nye county, approximately 90 miles northwest of Las Vegas in a sparsely populated, desert area. The impact analysis is focused on the communities that are within a fifty mile radius of the Yucca Mountain site. The study areas includes the unincorporated communities of Amargosa Valley, Beatty, and Pahrump in southern Nye county and Indian Springs in Clark county. Tax boundaries specified by the county commissioners are used to deliniate community boundaries for purposes of the impact analysis.

6. DATA AND METHODS

During a preliminary phase of the research, contacts were made with community leaders and residents. These contacts resulted in a network that later facilitated the collection of data. Field notes were taken describing the general layout of each community in the study area. These included the types and locations of businesses and residential areas. Four separate housing types were defined using the guidelines developed by the U.S. Bureau of the Census.

Following the preliminary investigation, the road system and other features were mapped for each community. Using these maps and utility records, representatives of the electrical company servicing southern Nye county identified the location and type of housing, if any, associated with all current electrical connections. This information was added to the housing unit file constructed from the utility records for each community. Because of the lack of adequate utility records for Indian Springs, housing information for this area was collected by a "windshield survey," a systematic, block-by-block canvassing of housing units by teams operating from automobiles (Lowe, Pittenger and Walker 1977). As a consequence, Indian Springs is not included in the test results reported in this paper.

The preliminary fieldwork indicated that substantial differences in *PPH* could be expected across the communities in the study area. Thus, a random selection of units from the housing unit file was drawn separately for each community, based on the number of housing units in each community. A conservative approach was used to determine the size of each community's sample. It assumed a 5% margin of error, a significance level of .05 and interest in a dichotomous variable with a 50-50 distribution (Cochran 1977). Once the initial size was determined, an additional 15% was added to allow for missing cases. The final sample size for Amargosa Valley was 175 housing units, for Beatty it was 222, and for Pahrump, 355.

Local experts were initially identified through the contact network on the basis of their experience in community activities and their familiarity with local residents. Each potential expert was interviewed and asked to complete a form designed to assess their qualifications. A written explanation of the project and specific instructions regarding the data collection

process were provided and discussed. The persons selected as local experts were given instructions regarding confidentiality. For this project, we found that the "meter readers" employed by the local utilities constituted a good source of local experts. The local experts were provided with the sample set of housing units for their respective communities. In most cases, two local experts worked together, which made it possible to verify the accuracy of information as it was recorded. For each unit, the local experts communicated to the researcher only the number of persons in the household as of July 15, 1990, the age (using eight age groups) and gender of each household member, and the retirement status of any member fifty years of age and over. If either of the two local experts was unsure about the composition of a given household, another member of the community was contacted to confirm the data. In the case where the composition of any part of the household could not be confirmed, "data unknown" was recorded for the entire unit. The data were recorded on a form that listed and identified the sample units by an attribute number (designated according to location on the housing unit map), the electrical meter number assigned to the unit, and the type of housing unit. All residential units, including those identified as "burned down" or otherwise destroyed, unoccupied or "removed from pad" (in the case of mobile homes and trailers) are considered part of the final sample. Units identified as "not a residence" were eliminated from the frame and not included in the sample. There were a few units for which data were unknown. These units are not included in the final sample, which may cause some slight bias.

7. RESULTS

The first data product is the number of households, which is derived directly from the active meter records, screened and classified by utility company personnel. Table 1 displays these figures by community along with other results that are discussed later.

Table 1 also provides the estimated *PPH*, which is taken from aggregate number of persons identified in the occupied sample units by the local experts. Also found in this table is the estimated household population of each community, which was found by applying the HUM formula to the household and *PPH* components. There were no group quarters identified in any of the communities.

Table 1
Sample Characteristics and Results of the Accuracy Test*

Community	Households	Estimated 1990			1990 Census Count	95% Confidence Interval	
		<i>PPH</i>	<i>SE</i>	Population		Low	High
Amargosa Valley	326	2.58	.11	841	853	771	911
Beatty	672	2.43	.10	1,633	1,623	1,501	1,765
Pahrump	3,224	2.23	.06	7,190	7,425	6,810	7,569

* The Estimated data and confidence intervals are produced by the procedures described in the text. The 1990 census counts are taken from Table 3 in the "1990 Census Extract, Nevada, Public Law 94-171 Data," dated February 11, 1991 and distributed by Betty McNeal, Nevada State Data Center Librarian, Nevada State Library and Archives, Capitol Complex, Carson City, Nevada 89710. The count for the area "Amargosa Valley is made up of the 1990 population reported for Nye county's Amargosa Valley Division (761) and Crystal Division (92). The count for the area "Beatty" is taken from the Beatty Census Designated Place and the count for the area "Pahrump" is taken from the Pahrump Division of Nye county.

8. MEASURING UNCERTAINTY IN THE ESTIMATES

One major advantage of estimates based on random sampling is that confidence intervals can be generated. Rives (1982) advocates this approach. However, he did not consider the use of local experts and believed that his suggestion would only be followed in exceptional circumstances because of the high expense associated with traditional surveys. This was also noted by Morrison (1982) and Lee and Goldsmith (1982) in their critical review of Rives' suggestion.

In the case of the local expert procedure, the "statistic" is the *PPH* value, which in practice would vary from sample to sample depending on the variation in *PPH* values. Our interest is less in the *PPH* values than in the estimate of population, however, so we use a simple transformation introduced by Espenshade and Tayman (1982) and used more recently by Swanson (1989) to place the confidence interval originally generated for a given community's *PPH* value around each of the community population estimates.

Let

$$\begin{aligned} P &= \text{estimated household population,} \\ N &= \text{number of households,} \\ PPH &= \text{estimated persons per household.} \end{aligned}$$

Then

$$\begin{aligned} \text{lower limit } (P) &= (N) * (PPH - (t_{n-2}, \alpha/2) * (se)), \\ \text{upper limit } (P) &= (N) * (PPH + (t_{n-2}, \alpha/2) * (se)), \end{aligned}$$

where

$$\begin{aligned} n &= \text{number of households sampled,} \\ \alpha &= \text{level of significance desired,} \\ se &= \text{standard error of the estimated } PPH, \\ t_{n-2} &= (\alpha/2)100\text{th percentile of the } t \text{ distribution, with } (n - 2) \text{ degrees of freedom.} \end{aligned}$$

As an example, using a significance level of .05, the corresponding 95% confidence interval for the estimated 1990 population of Pahrump (7,190) is

$$\begin{aligned} \text{lower limit} &= 6,810 = (3,224) * (2.23 - (1.96 * .06)), \\ \text{upper limit} &= 7,569 = (3,224) * (2.23 + (1.96 * .06)). \end{aligned}$$

9. TEST OF ACCURACY

Before turning to the test results, which are also included in Table 1, some data qualifications require discussion. The single most problematic issue in terms of comparing the estimates with the 1990 census results lies in the fact that the Bureau of the Census does not recognize the "tax districts" as administrative boundaries for the communities in the study area. This means that the Bureau's "statistical" geography must be used, which requires some adjustments so that the geography used for purposes of the impact analysis matches that used by the Bureau.

In terms of these adjustments, the area identified as Amargosa Valley for purposes of the impact study is known to vary from the Amargosa Valley Census Division of Nye county used by the Bureau in that the study's definition includes the Crystal Census Division of Nye county. Fortunately, this is a case where two pieces of statistical geography used by the Bureau can be combined to virtually match that used in the impact study. Thus, the 1990 census population counts shown in Table 1 for the Amargosa Valley include the Crystal Division along with the Amargosa Valley Division. "Beatty" is another area that is known to vary in terms of geography. It is identified as both a Census Designated Place and as the Beatty Census Division of Nye county by the Bureau. In this situation, it is the Census Designated Place that corresponds very closely to the definition of Beatty used in the impact study. Thus, the 1990 census population count for Beatty shown in Table 1 is for the Beatty CDP.

The third community, Pahrump, is identified as a Census Division of Nye county. This piece of statistical geography used by the Bureau is virtually identical to that used in the impact study. Consequently, the 1990 census population found in Table 1 for Pahrump is that given for this division of Nye county.

There are other differences between the estimates and the 1990 census figures. The official date of the census count is April 1st; the estimates are for July 15th. In terms of this difference, seasonal effects are believed to be very slight for the communities in question. With the exception of the outflow of some "snowbirds," who may have been counted in the study area because they had no usual residence elsewhere, there were no known migration streams of any consequence between April and July. Similarly, the other components of population change were slight.

Had the Bureau found transient persons with no usual residence elsewhere, the estimation procedure is likely to have missed them. These differences would also impact housing unit counts. If a transient person, identified as a resident for purposes of the decennial census, is found in a recreational vehicle it would be included in the community's "other" housing stock by the Bureau. Such accommodations would not be included in the data derived from the residential electrical meter records.

We believe, however, that such instances are rare and, further, that the test results are not confounded by comparing a household population with a population that resides mainly in households but also, to some extent, in group quarters.

The results of the test of accuracy are also summarized in Table 1, along with the "low" and "high" estimates corresponding to the 95 percent confidence interval placed around each community's estimated population. The estimated population is very close to the population reported by the Bureau. Overall, the mean absolute difference is 86 persons and the mean absolute percent difference is 1.7.

The three confidence intervals contain the 1990 census population in each of the three communities, respectively. This finding is of special interest because the intervals are relatively narrow for a 95 percent level of confidence. On average, the width, as measured from the estimated population to either boundary is 7.2 percent of the estimated population. This suggests that confidence intervals constructed around the estimates derived from this variation of the HUM are meaningful, even in the presence of some unknown level of nonsampling error.

Two of the three communities are underestimated. In the case of Pahrump, it appears that the estimation technique was not able to capture all of the recent growth that appears to be spilling from the Las Vegas Valley into the Pahrump area. It is not known to what extent this was due to missing households on the frame and what was due to underestimating Pahrump's *PPH* value.

10. SUMMARY

While the local expert procedure may not provide satisfactory population estimates in all small, rural areas (*e.g.*, vacation spots, with a high incidence of seasonal housing units and privately owned rental units), it appears to hold promise based on the data for the area included in this study. As with any estimation technique, the key criteria for determining if it could be implemented elsewhere revolve around the possibility of obtaining the required data and implementing the procedure within available funding. In the case of the local expert procedure, this would mean that utility data can provide the number of households and be used as a sample frame. Once a sample was selected, the procedure's effectiveness would depend on the recruitment and knowledge of local experts. If these criteria can be met, the procedure would seem to be feasible. The next step would be to determine how accurate it is in a given application.

We were not able to evaluate the accuracy of the age and other composition data estimated through the procedure at the time of this writing because these data were not yet available from the 1990 decennial census. However, we are encouraged by the test results for the total population, which indicate that the procedure has the potential for highly accurate estimates, even in small, rural areas experiencing rapid change.

ACKNOWLEDGMENT

The authors are grateful for comments made by Mark Flotow and anonymous reviewers on earlier drafts of this paper.

REFERENCES

- BROWN, R., GEERTSEN, H., and KRANNICH, R. (1989). Community satisfaction and social integration in a boomtown: A longitudinal analysis. *Rural Sociology*, 54, 568-586.
- BYERLY, E. (1990). State and local agencies preparing population and housing estimates. *Current Population Reports Series P-25*, 1063. Washington, DC: U.S. Bureau of the Census.
- CARLSON, J., SWANSON, D., and WILLIAMS, C. (1990). The development of small area socioeconomic data to be utilized for impact analysis: Rural, Southern Nevada. In *High Level Radioactive Waste Management Proceedings of the 1990 International Conference*. LaGrange Park, Illinois: American Nuclear Society, 985-990.
- COCHRAN, W. (1977). *Sampling Techniques*, (3rd Edition). New York: Wiley.
- ESPENSHADE, T., and TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.
- FREUDENBURG, W. (1982). Social impact assessment. In *Rural Society in the U.S.: Issues for the 1980s*, (Eds. D. Dillman and D. Hobbs). Boulder, Colorado: Westview Press, 296-303.
- KRANNICH, R., BERRY, E., and GREDIER, T. (1989). Fear of crime in rapidly changing rural communities: A longitudinal analysis. *Rural Sociology*, 54, 195-212.
- KRANNICH, R., and GREDIER, T. (1984). Personal well-being in rapid growth and stable communities: Multiple indicators and contrasting results. *Rural Sociology*, 49, 541-552.
- LEE, E., and GOLDSMITH, H. (1982). Evaluation and synopsis. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 113-118.
- LOWE, T., PITTENGER, D., and WALKER, J. (1977). Making the housing unit method work: A progress report. Presented at the Annual Meeting of the Population Association of America.

- LOWE, T., WEISSER, L., and MYERS, B. (1984). A special consideration in improving housing unit method estimates: The interaction effect. Presented at the Annual Meeting of the Population Association of America.
- MARTIN, J., and SEROW, W. (1979). Virginia's state-local cooperative program: conflict and cooperation in producing population estimates. *State Government*, 182-186.
- MORRISON, P. (1982). Alternatives for monitoring local demographic change. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 97-111.
- MUNSELL, K. (1988). Towns cope with hazardous wastes. *Small Town*, 18, 30.
- PITTENGER, D. (1976). *Projecting State and Local Populations*. Cambridge, Massachusetts: Ballinger Press.
- POGGIE, J. (1972). Toward quality control in key informant data. *Human Organization*, 31, 23-30.
- RIVES, N. (1982). Assessment of a survey approach. In *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 79-96.
- RIVES, N., and SEROW, W. (1984). *Introduction to Applied Demography Data Sources and Estimation Techniques*. Beverly Hills, California: Sage.
- SCHLEIFER, S. (1986). Trends in attitudes toward and participation in survey research. *Public Opinion Quarterly*, 50, 17-26.
- SEIDLER, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organizational analysis. *American Sociological Review*, 39, 816-831.
- SMITH, S. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, 81, 287-296.
- SMITH, S., and LEWIS, B. (1983). Some new techniques for applying the housing unit method of local population estimation: Further evidence. *Demography*, 20, 407-413.
- SMITH, S., and LEWIS, B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography*, 17, 323-339.
- SMITH, S., and MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 79, 282-289.
- STARSINIC, D., and ZITTER, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. *Demography*, 5, 475-484.
- SUMMERS, G. (1982). *Industrialization*. In *Rural Society in the U.S.: Issues for the 1980s*, (Eds. D. Dillman and D. Hobbs). Boulder, Colorado: Westview Press, 164-174.
- SWANSON, D. (1989). Confidence intervals for postcensal population estimates: A case study for local areas. *Survey Methodology*, 15, 271-280.
- SWANSON, D., BAKER, B., and VAN PATTEN, J. (1983). Conceptual and practical features of the housing unit method. Presented at the Annual Meeting of the Population Association of America.
- U.S. BUREAU OF THE CENSUS (1978). State and local agencies preparing population estimates and projections: Survey of 1975-76. *Current Population Reports P-25*, 725. Washington, DC: U.S. Bureau of the Census.
- U.S. DEPARTMENT OF ENERGY (1988). Section 175 report. *Secretary of Energy's Report to the Congress Pursuant to Section 175 of the Nuclear Waste Policy Act, As Amended*. Washington, DC: Office of Civilian Radioactive Waste Management, U.S. Department of Energy.

Single Stage Cluster Sampling in Prevalence-Incidence Surveys: Some Issues Suggested by the Shanghai Survey of Alzheimer's Disease and Dementia

ZHISEN XIA, PAUL S. LEVY, ELENA S.H. YU, ZHENGYU WANG,
and MINGYUAN ZHANG¹

ABSTRACT

The scenario considered here is that of a sample survey having the following two major objectives: (1) identification for future follow up studies of n^* subjects in each of H subdomains, and (2) estimation as of this time of conduct of the survey of the level of some characteristic in each of these subdomains. An additional constraint imposed here is that the sample design is restricted to single stage cluster sampling. A variation of single stage cluster sampling called *telescopic single stage cluster sampling* (TSSCS) had been proposed in an earlier paper (Levy *et al.* 1989) as a cost effective method of identifying n^* individuals in each sub domain and, in this article, we investigate the statistical properties of TSSCS in crosssectional estimation of the level of a population characteristic. In particular, TSSCS is compared to ordinary single stage cluster sampling (OSSCS) with respect to the reliability of estimates at fixed cost. Motivation for this investigation comes from problems faced during the statistical design of the Shanghai Survey of Alzheimer's Disease and Dementia (SSADD), an epidemiological study of the prevalence and incidence of Alzheimer's disease and dementia.

KEY WORDS: Single stage cluster sampling; Prevalence estimation; Telescopic single stage cluster sampling; Alzheimer's disease; Dementia.

1. BACKGROUND AND INTRODUCTION

Many studies have both a crosssectional component in which the levels of quantitative variables or prevalences of dichotomous variables are estimated by means of a sample survey, and a longitudinal component in which a cohort of individuals is identified by means of the same sample survey and followed over a defined period for subsequent events. This type of study is especially common in the field of epidemiology in which estimates of the prevalence of a disease or condition are required both for the study population as a whole as well as for defined subgroups of it, and a sufficient number of individuals initially free of the disease or condition need to be identified within each of the defined subgroups for future estimation of the incidence of the disease or condition (*cf.* Kannel 1966).

Design of a cost efficient sampling plan for such studies poses a challenge since sufficient numbers of individuals within each domain must be selected, often under some type of cluster sampling scheme, to ensure reliable estimation of both the prevalence and incidences discussed above. In this report, which has been motivated by a recent study conducted in China, we discuss these issues of sample design under a particular type of cluster sampling (single stage cluster sampling).

¹ Zhisen Xia, Paul S. Levy and Zhengyu Wang, School of Public Health, University of Illinois at Chicago, P.O. Box 6998, Chicago IL. 60680, U.S.A., Elena S.H. Yu, School of Public Health, San Diego State University, San Diego, California, U.S.A., Mingyuan Zhang, Shanghai Institute of Mental Health, Shanghai, Peoples Republic of China.

2. STATISTICAL FORMULATION

Let us suppose that a population consists of N individuals divided into H mutually exclusive subdomains, each containing N_h individuals ($h = 1, \dots, H$). Suppose further that the population is grouped into M clusters which will comprise the sampling units for the survey. Let us assume that sampling of the clusters will be according to ordinary single-stage cluster sampling (*i.e.*, simple random sampling of clusters followed by selection of all individuals within each sample cluster.)

If we wish to identify with $100 \times (1 - \alpha)\%$ confidence at least n_h^* individuals in a particular domain, h , then the following number, m'_h , of clusters must be selected (*cf.* Levy *et al.* 1989):

$$m'_h = \left[A_h + \left(A_h^2 + \frac{n_h^*}{\bar{N}_h} \right)^{1/2} \right]^2, \tag{1}$$

where

N_{hi} = the number of individuals in domain h , cluster i , ($i = 1, \dots, M$),

$$\bar{N}_h = \sum_{i=1}^M N_{hi}/M,$$

$$V_{N_h} = \sigma_{N_h}/\bar{N}_h,$$

$$\sigma_{N_h}^2 = \sum_{i=1}^M (N_{hi} - \bar{N}_h)^2/(M - 1),$$

z_α = the 100α 'th percentile of the normal distribution

and

$$A_h = |z_\alpha| \times V_{N_h}/2.$$

The above assumes that the N_{hi} are normally distributed over the M clusters. Also, the number, n_h^* , of individuals needed in domain h is based on statistical considerations relevant to the longitudinal component of the study. For example, it could be based on the expected occurrence rate of the event of interest in the follow up period and the precision required for the estimate of this occurrence rate.

If one also wishes to estimate with $100 \times (1 - \alpha)\%$ confidence the total or mean level of some variable \mathcal{H} to within $100 \times \epsilon\%$ of its true value for each domain, h , then one would require sampling of the following number, M''_h , of clusters in domain h ;

$$m''_h = \frac{z_{1-\alpha/2}^2 M V_{hx}^2}{z_{1-\alpha/2}^2 V_{hx}^2 + (M - 1)\epsilon^2}, \tag{2}$$

where,

X_{hij} = the level of variable \mathcal{X} for individual j within domain h of cluster i
 $(j = 1, \dots, N_{hi}; i = 1, \dots, M),$

$$X_{hi} = \sum_{j=1}^{N_{hi}} X_{hij},$$

$$\bar{X}_h = \sum_{i=1}^M X_{hi}/M,$$

$$\sigma_{hx}^2 = \sum_{i=1}^M (X_{hi} - \bar{X}_h)^2/M,$$

and

$$V_{hx}^2 = \sigma_{hx}^2 / \bar{X}_h^2.$$

For both of the specifications stated above to be satisfied within each domain, it follows that we would require m_h clusters to be sampled where for $h = 1, \dots, H,$

$$m_h = \max(m'_h, m''_h). \tag{3}$$

Without loss of generality, we can relabel the domains in order of increasing required m_h (i.e., $m_1 \leq m_2 \leq \dots \leq m_H$).

Finally, in order for both of the specifications to be satisfied in each of the H domains under an ordinary single stage cluster sampling design, the number, m , of clusters required to be sampled would be m_H . We note again that in ordinary single stage cluster sampling, every individual in every sample cluster is sampled. Thus, while the specifications of sample size are met minimally in domain H , the domain requiring the largest number of sample clusters, they are more than met in the other domains: $1, \dots, H - 1$. This inclusion in domains other than H of more individuals than are actually required could result in a survey that has overly expensive field costs.

The alternative to ordinary single-stage cluster sampling that is generally used to avoid this needless expense would be a two-stage cluster sampling design with different second stage sampling fractions (i.e., over sampling) in each domain. Given, however, a scenario in which it is not feasible to subsample at all within clusters, a methodology called *single stage telescopic cluster sampling* (SSTCS) was proposed in an earlier publication (Levy *et al.* 1989) which allowed the *eligibility rule* (i.e., the rule that determines which individuals are eligible for inclusion in the sample) to vary over the sample clusters. In this design, the particular domains included in the sample would not be the same for each sample cluster. This earlier publication demonstrated the usefulness of single stage telescopic sampling in surveys which have as major objective the identification for future longitudinal follow up of a certain number of individuals in each of several domains. In this report, we will characterize the properties of estimates from this type of design and compare them to estimates from ordinary single-stage cluster sampling.

3. TELESCOPIC SINGLE-STAGE CLUSTER SAMPLING

3.1 Sampling of Clusters

As mentioned above, single-stage telescopic cluster sampling is proposed as a cost saving alternative to ordinary single-stage cluster sampling in situations where it is not feasible to sub-sample within sample clusters, and is performed as follows. If there are H mutually exclusive and exhaustive domains for which estimates are desired, and if m clusters are to be sampled, the m sample clusters are divided randomly into m_1^* type 1 clusters, m_2^* type 2 clusters, \dots , and m_H^* type H clusters having the following properties: A type h cluster ($h = 1, \dots, H$) as illustrated below has as eligible sample persons individuals in domains $h, h + 1, \dots, H$, but not in domains h' where $h' < h$.

Cluster Type	Domains Sampled			
	1	2	h	H
1	+	+	+	+
2	−	+	+	+
h	−	−	+	+
H	−	−	−	+

“+” = domain sampled

“−” = domain not sampled.

The term *telescopic* was suggested by the appearance of the above diagram.

The number, m_h^* , of type h clusters is generally determined according to the following strategies: Suppose that under single-stage cluster sampling, a sample of m_h clusters as determined by relation (3) is required for domain h , ($h = 1, \dots, H$); and, again supposing that $m_1 \leq m_2 \dots \leq m_H$, we would let:

$$m_1^* = m_1; \text{ and } m_h^* = m_h - m_{h-1} \text{ for } h = 2, \dots, H.$$

Clearly, this allocation results in a total of m_H sample clusters being selected, with elements in each domain, h , being sampled in m_h sample clusters, exactly the number of clusters required to achieve the specifications placed on the reliability of estimates and the identification of individuals for future follow up. As discussed above, if ordinary single-stage cluster sampling (OSSCS) were used, a sample of m_H clusters would be needed to meet specifications in domain H , but this would entail individuals in the other domains also being sampled in m_H clusters in excess of that needed to meet the stated specifications.

3.2 Characterization of Estimates

Let

$$\sigma_{h k x} = \sum_{i=1}^M (X_{hi} - \bar{X}_h)(X_{ki} - \bar{X}_k)/M,$$

$S_h = \{i_1, i_2, \dots, i_{m_h}\}$ = the set of sample clusters having eligible persons in domain h .

The following results can then be obtained from combinatorial theory.

1. The estimated total, x'_{tel} , under TSSCS of a population total X is given by

$$x'_{tel} = \sum_{h=1}^H x'_h, \quad (4)$$

where x'_h is given by

$$x'_h = (M/m_h) \sum_{i \in S_h} X_{hi}.$$

2. The mean, $E(x'_{tel})$, and variance, $\text{Var}(x'_{tel})$, of x'_{tel} are given by

$$E(x'_{tel}) = X, \quad (5)$$

$$\text{Var}(x'_{tel}) = \sum_{h=1}^H \frac{M^2}{m_h} \left(\frac{M - m_h}{M - 1} \right) \left(\sigma_{hx}^2 + 2 \sum_{k < h} \sigma_{h k x} \right). \quad (6)$$

These relationships follow in a straightforward way from combinatorial theory.

4. COST COMPARISONS BETWEEN OSSCS AND TSSCS

We can examine the comparative costs of OSSCS vs. TSSS by considering the following simple cost function that would be associated with OSSCS:

$$C_0 = C_1 m_H + C_2 m_H (\bar{N}_1 + \bar{N}_2 + \dots + \bar{N}_H) = m_H \left(C_1 + C_2 \sum_{h=1}^H \bar{N}_h \right), \quad (7)$$

where C_0 is the expected cost, C_1 is the cost component associated with clusters (*e.g.*, travel to and from cluster, procurement of the list of enumeration units in the cluster, preparation of materials for field work within the cluster, *etc.*) and C_2 is the cost component associated with listing units (primarily travel between listing units and interviewing). It should also be noted that the expression, $\sum_{h=1}^H \bar{N}_h$, is the average number of listing units per cluster. Again, throughout this discussion the listing units are the individuals themselves. The analogous expected cost, C_t , associated with telescopic sampling would then be given by:

$$C_t = C_1 m_H + C_2 (m_1 \bar{N}_1 + m_2 \bar{N}_2 + \dots + m_H \bar{N}_H) = m_H \left(C_1 + C_2 \sum_{h=1}^H \gamma_h \bar{N}_h \right), \quad (8)$$

where $\gamma_h = m_h/m_H$ (which is ≤ 1). Thus, the cost, C_t , associated with TSSCS is less than or equal to that associated with an OSSCS of the same number of clusters with the difference being equal to

$$C_2 m_H \sum_{h=1}^H (1 - \gamma_h) \bar{N}_h.$$

The most important comparison between the two sample designs, in many instances, would be that involving their performance at equivalent cost in estimating the overall level, X , of a characteristic, \mathcal{Y} . An estimator, x'_{ord} , based on an OSSCS of m_H clusters (the number required to meet the specifications within each domain) would have variance given by:

$$\text{Var}(x'_{ord}) = \frac{M^2}{m_H} \left(\frac{M - m_H}{M - 1} \right) \sum_{h=1}^H \left(\sigma_{hx}^2 + 2 \sum_{k < h} \sigma_{h k x} \right). \tag{9}$$

This is not the usual form of the variance (*cf.* Levy and Lemeshow 1991, chapter 9), but is an algebraically equivalent form that can be compared directly with the variance of x'_{tel} based on a TSSCS design with m_H clusters sampled (equation (6)). The difference between these two variances is given by

$$\text{Var}(x'_{tel}) - \text{Var}(x'_{ord}) = \frac{M^3}{M - 1} \sum_{h=1}^H \left(\frac{m_H - m_h}{m_H m_h} \right) \sum_{h=1}^H \left(\sigma_{hx}^2 + 2 \sum_{k < h} \sigma_{h k x} \right), \tag{10}$$

which is greater than or equal to zero (0). This is not surprising since an OSSCS of m_H clusters will invariably result in a larger overall sample size than a TSSCS of the same number of clusters.

Although an OSSCS of m_H clusters will result in an estimator, x'_{ord} , which has a lower variance than the estimator, x'_{tel} , resulting from a TSSCS of the same number, m_H , of clusters, it does so at a higher cost. For this reason, it is more reasonable to compare x'_{tel} based on a sample of m_H clusters to x'_{ord} based on a sample of m^* clusters where m^* is the number of clusters that can be sampled from an OSSCS design at cost equivalent to that based on a TSSCS design having m_H sample clusters. From equations (7) and (8), it follows that m^* is given by:

$$m^* = m_H \left(\frac{1 + \frac{C_2}{C_1} \sum_{h=1}^H \gamma_h \bar{N}_h}{1 + \frac{C_2}{C_1} \sum_{h=1}^H \bar{N}_h} \right). \tag{11}$$

It should be noted that

- (1) $m^* \leq m_H$.
- (2) As $C_2/C_1 \rightarrow \infty$, then $m^* \rightarrow \bar{m}_w$

where,

$$\bar{m}_w = \sum_{h=1}^H m_h \bar{N}_h \bigg/ \sum_{h=1}^H \bar{N}_h.$$

- (3) As $C_2 \backslash C_1 \rightarrow 0$, then $m^* \rightarrow m_H$

and

- (4) m^* decreases monotonically with increase in C_2/C_1 which implies that $\bar{m}_w \leq m^* \leq m_H$.

From the above analysis, we note that at a cost equivalent to that of a TSSCS of m_H clusters, the variance of x'_{ord} (ignoring the finite population correction) will be inflated by at most a factor equal to m_H/\bar{m}_w over that which would have been obtained from an OSSCS of m_H clusters, where \bar{m}_w is a weighted mean of the m_h clusters required within each domain for the domain specific specifications to be met. The weights in this instance are the \bar{N}_h , which are the average number of individuals within each particular domain. It should be noted also that the reduction in effective sample size of an OSSCS equivalent in cost to a TSSCS increases with increase in C_2/C_1 , which is essentially the ratio of the cost of extracting information from sample individuals to that of preparing the sample clusters for the survey. This makes sense intuitively.

The issues discussed above are illustrated in the next section with data from the Shanghai Survey of Alzheimer's Disease and Dementia.

5. THE SHANGHAI SURVEY OF ALZHEIMER'S DISEASE AND DEMENTIA

The SSADD was planned in 1986 having as major objectives: (1) estimation of the prevalence of physical and mental impairments including Alzheimer's and other dementing diseases among persons in each of three age groups (55-64 yrs/65-74 years/ and 75 yrs. and older) in the Jing-An district of Shanghai, China, and (2) identification of approximately 1,400 persons in each of these 3 age groups for future determination of the incidence of these conditions. Jing-An, is one of twelve districts comprising the city of Shanghai, and was chosen as the target area because of its relatively large and stable population of elderly and its proximity to the Shanghai Institute of Mental Health which was responsible for the field work. Findings from this study have been discussed by Zhang *et al.* (1990) and by Yu *et al.* (1989). Methodological issues have been discussed by Levy *et al.* (1988 and 1989).

The clusters in this survey are administrative entities called *neighborhood groups* consisting of geographically contiguous households having a well defined social and political structure. The strategy was to involve the leaders of neighborhood groups selected in the sample in the identification and recruitment of eligible persons. At the time of the planning of the survey, there were 4,066 neighborhood groups within the Jing-An District. This particular population of aging and elderly Chinese generally had a low level of education and had experienced in their lifetimes repeated periods of political upheaval and repression (*e.g.*, the Warlords, the Japanese invasion, the Cultural Revolution), where being singled out or selected often had adverse consequences. For these reasons, it was felt strongly, especially by the local Chinese members of the research team who were most familiar with the target population, that any attempt to subsample persons in the target age groups within neighborhood groups that fall into the sample would compromise response rates and overall cooperation.

Restricted to single stage cluster sampling and faced with a very tight deadline for designing the sample, the member of the study team responsible for the sample design (PSL) proposed a heuristic method that would result with reasonable certainty in the identification of 1,400 individuals within each of the three target age groups. The resulting design was essentially a TSSCS in which 446 neighborhood groups were sampled. For details of this design, the reader is referred to the publications on the SSADD cited above. It should be emphasized that the resulting design was chosen purely on heuristic grounds and long before the theory behind this methodology was developed.

Of the 446 neighborhood groups sampled, 149 were designated as type 1, and 136 of these contained at least 1 person in the target age group (55 years and above). Since only the type 1 clusters have as eligible respondents all persons in each of the 3 target age groups, they can be used to estimate all of the parameters needed to evaluate the cost effectiveness of TSSCS relative to OSSCS. In the ensuing discussion, we will use the data from these 136 clusters to illustrate numerically how, on the basis of available “pilot” data, comparisons can be made between OSSCS and TSSCS with respect to cost effectiveness. From this sample of 136 clusters, we have for each domain, h , estimates of relevant parameters as shown below:

Age	\bar{N}_h	V_{N_h}	\bar{X}_h	V_{hx}	$\sum_{k < h} \sigma_{h k x}$
55-64	10.985	.485	.125	2.991	0.000
65-74	8.088	.513	.360	2.357	0.190
75 +	3.478	.643	.456	1.665	0.296.

If we wish to identify with 95% confidence at least 1,400 persons in each age group, then from relation (1) and the data shown above, we would have

$$A_1 = 1.645 \times 0.485/2 = 0.3989$$

and

$$m'_1 = \left[0.3989 + \left((0.3989)^2 + \frac{1,400}{10.985} \right)^{1/2} \right]^2 = 136.78 \approx 137.$$

Similarly, $m'_2 \approx 185$, and $m'_3 \approx 419$.

Let us suppose that for each of the three age groups, we wish to estimate with 80% confidence to within 30% of its true value the proportion, \bar{X}_h , of persons showing evidence of cognitive dysfunction as judged by a score below 18 on the Mini Mental State Examination (MMSE), which is a screening test for cognitive dysfunction. From these same data, we have the following estimates of the parameters necessary to determine the number of sample clusters required to meet this specification:

From relation (2), with $M = 4,066$, $\epsilon = 0.30$, and $z_{1-\alpha/2} = 1.28$, we have the following values of m''_h :

$$m''_1 = 157; \quad m''_2 = 99; \quad m''_3 = 50$$

and from relation (3), the number, m_h , of clusters required to satisfy both conditions in each domain is given by:

$$m_1 = \max(137, 157) = 157; \quad m_2 = \max(185, 99) = 185; \quad m_3 = \max(419, 50) = 419.$$

Thus, for an OSSCS design to satisfy both specifications, the number, m , of clusters required to be sampled would be 419. Likewise, a TSSCS design having 157 type 1, 28 type 2, and 234 type 3 sample clusters would satisfy both requirements.

The cost components, C_1 and C_2 , expressed in person hours, are estimated to be 20 and 2 respectively. The relatively high cost component, C_1 , associated with clusters is due to the fact that once a neighborhood group is selected in the sample, many hours must be spent obtaining the list of households and persons from a central bureau and enlisting the support of the neighborhood group leaders. The cost component, C_2 , of 2 person hours associated with individuals involves primarily interview and call-back activities. Thus, the field costs, C_0 , associated with an OSSCS design that satisfies both specifications is (from relation (7)) 27,278 person hours as compared to a cost of 17,737 person hours (from relation (8)) associated with a TSSCS design that satisfies both specifications. This represents a 35% savings in field costs, which is substantial.

From relation (9), we calculate that the estimate, x'_{ord} , of the number of persons over all 3 age groups having evidence of a cognitive disorder based on an OSSCS of 419 sample clusters would have variance equal to 70,844, whereas x'_{tel} , the analogous estimate based on a TSSCS also with 419 clusters, would have variance equal to 122,744, which is 42% greater than the variance of the OSSCS estimate. However, an OSSCS design having the same field costs as a TSSCS design based on 419 sample clusters would permit only 208 clusters to be sampled (relation (11)). The variance of x'_{ord} based on an OSSCS design with 208 sample clusters would be estimated to be 141,733, which is 15% higher than the variance of the analogous TSSCS estimate having the same field cost. Also, the OSSCS design having 208 sample clusters would not satisfy the two specifications placed on the estimates.

6. DISCUSSION

The methodology, TSSCS, discussed here and in earlier publications, arose from a situation in which cluster sampling was clearly indicated but a definite "red light" was given to any subsampling within clusters. For the Shanghai Survey of Alzheimer's Disease and Dementia considered here, the two major objectives were to identify a certain number of individuals within each of 3 domains (age groups in this instance) and to obtain domain specific estimates meeting certain specifications pertaining to precision. Based on results presented above for this particular survey, it appears that this method could result in considerable savings in field costs without compromising objectives.

One might raise questions concerning the general applicability of this methodology. It would be of use primarily in situations where it is either not feasible or too costly to subsample clusters and the individuals do not have to be screened to determine whether they belong to one of the target domains (in the SSADD, the leadership of the sample neighborhood groups provided a list of all persons in the neighborhood group along with information on data of birth). Such scenarios may occur, for example, in surveys where data are abstracted from records by personnel sufficiently familiar with the records to abstract information, but not considered capable of sampling the records without expensive supervision. Again, in such situations, TSSCS may provide a reasonable alternative.

ACKNOWLEDGEMENTS

Research on this report was supported, in part, by grant number 1 RO1 AG10327-01 from the National Institute on Aging.

REFERENCES

- KANNEL, W.B. (1966). An epidemiological study of cerebrovascular disease. In *Fifth Conference on Cerebrovascular Diseases*, (Eds. R.G. Sickert and J.P. Whisnatt). New York: Grune and Stratton.
- LEVY, P.S., YU, E.S.H., LIU, W.T., ZHANG, M., WANG, Z., WONG, S., and KATZMAN, R. (1988). Variation on single stage cluster sampling used in a survey of elderly people in Shanghai. *International Journal of Epidemiology*, 17, 931-933.
- LEVY, P.S., YU, E.S.H., LIU, W.T., WONG, S., ZHANG, M., WANG, Z., and KATZMAN, R. (1989). Single-stage cluster sampling with a telescopic respondent rule: A variation motivated by a survey of dementia in elderly residents of Shanghai. *Statistics in Medicine*, 8, 1537-1544.
- LEVY, P.S., and LEMESHOW, S. (1991). *Sampling of Populations: Methods and Applications*. New York: John Wiley and Sons.
- YU, E.S.H., LIU, W.T., LEVY, P.S., ZHANG, M., KATZMAN, R., LUNG, C.T., WONG, S., and WANG, Z. (1989). Cognitive impairment among elderly in Shanghai, China. *Journal of Gerontology Social Sciences*, 44, 97-106.
- ZHANG, M., KATZMAN, R., SALMON, D., JIN, H., CAI, G., WANG, Z., QU, G., GRANT, I., YU, E.S.H., LEVY, P.S., KLAUBER, M.R., and LIU, W.T. (1990). The prevalence of dementia and Alzheimer's disease in Shanghai China: Impact of age, gender, and education. *Annals of Neurology*, 27, 428-437.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents JOS 1991, Volume 7, Number 4

A Framework for Analyzing Categorical Survey Data with Nonresponse <i>David A. Binder</i>	393
Inference with Survey Weights <i>Roderick J.A. Little</i>	405
Cluster Analysis in International Comparisons <i>György Szilágyi</i>	425
A Multivariate Time Series Analysis of Fertility, Adult Mortality, and Real Wages in Sweden 1751-1850: A Comparison of Two Different Approaches <i>Mats Hagnell</i>	437
Census by Questionnaire - Census by Registers and Administrative Records: The Experience of Finland <i>Pekka Myrskylä</i>	457
A Comparison of the Missing-Data Treatments in the Post-Enumeration Program <i>Joseph L. Schafer</i>	475
Miscellanea	
Natural Resource and Environmental Accounting in the National Accounts <i>Lindy H. Ingham</i>	499
In Other Journals	515
Special Notes	519
Book Reviews	521
Index to Volume 7, 1991	531

All inquiries about submissions and subscriptions should be directed to the Chief Editor:

Lars Lyberg, U/SFI, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 41, No. 2, 1992

	<i>Page</i>
Analysis of ordinal paired comparison data <i>A. Agresti</i>	287
Robust procedures for drug combination problems with quantal responses <i>T. J. Vidmar, J. W. McKean and T. P. Hettmansperger</i>	299
Pseudofactors: normal use to improve design and facilitate analysis <i>H. Monod and R. A. Bailey</i>	317
Adaptive rejection sampling for Gibbs sampling <i>W. R. Gilks and P. Wild</i>	337
A note on Wadley's problem with overdispersion <i>B. J. T. Morgan and D. M. Smith</i>	349
Sequential application of Wilks's multivariate outlier test <i>C. Caroni and P. Prescott</i>	355
Influence in correspondence analysis <i>P. Pack and I. T. Jolliffe</i>	365
A reduced rank regression model for local variation in solar radiation <i>C. A. Glasbey</i>	381
Hierarchical Bayesian analysis of changepoint problems <i>B. P. Carlin, A. E. Gelfand and A. F. M. Smith</i>	389
<i>General Interest Section</i>	
Interlaboratory comparisons: round robins with random effects <i>M. J. Crowder</i>	409
<i>Letter to the Editors</i>	427
<i>Book Reviews</i>	429
<i>Statistical Software Reviews</i>	
DYMO	436
SPSS/PC+	438
<i>Statistical Algorithms</i>	
AS 273 Comparing subsets of regressor variables <i>A. J. Miller</i>	443
AS 274 Least squares routines to supplement those of Gentleman <i>A. J. Miller</i>	458
AS 275 Computing the non-central χ^2 distribution function <i>C. G. Ding</i>	478
AS 276 Normal combinatoric classification <i>C. L. Dunn</i>	483
<i>Remark</i>	
AS R89 A remark on Algorithm AS 76: An integral useful in calculating non-central t and bivariate normal probabilities <i>P. W. Goedhart and M. J. W. Jansen</i>	496

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

- 1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
- 2. **Résumé**
 - Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

- 3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(.) et log(.) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.
- 4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)
- 5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
 - 5.2 Exemple: Cochran (1977, p. 164). La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

BIBLIOGRAPHIE

- KANNEL, W.B. (1966). An epidemiological study of cerebrovascular disease. Dans *Fifth Conference on Cerebrovascular Diseases*, (Eds. R.G. Sickett et J.P. Whisnart). New York: Grune and Stratton.
- LEVY, P.S., YU, E.S.H., LIU, W.T., ZHANG, M., WANG, Z., WONG, S., et KATZMAN, R. (1988). Variation on single stage cluster sampling used in a survey of elderly people in Shanghai. *International Journal of Epidemiology*, 17, 931-933.
- LEVY, P.S., YU, E.S.H., LIU, W.T., WONG, S., ZHANG, M., WANG, Z., et KATZMAN, R. (1989). Single-stage cluster sampling with a telescopic respondent rule: A variation motivated by a survey of dementia in elderly residents of Shanghai. *Statistics in Medicine*, 8, 1537-1544.
- LEVY, P.S., et LEMESHOW, S. (1991). *Sampling of Populations: Methods and Applications*. New York: John Wiley and Sons.
- YU, E.S.H., LIU, W.T., LEVY, P.S., ZHANG, M., KATZMAN, R., LUNG, C.T., WONG, S., et WANG, Z. (1989). Cognitive impairment among elderly in Shanghai, China. *Journal of Gerontology Social Sciences*, 44, 97-106.
- ZHANG, M., KATZMAN, R., SALMON, D., JIN, H., CAI, G., WANG, Z., QU, G., GRANT, I., YU, E.S.H., LEVY, P.S., KLAUBER, M.R., et LIU, W.T. (1990). The prevalence of dementia and Alzheimer's disease in Shanghai China: Impact of age, gender, and education. *Annals of Neurology*, 27, 428-437.

Les éléments C_1 et C_2 du coût total C_0 , exprimés en heures-personne, sont estimés à 20 et à 2 respectivement. Le coût C_1 associé aux grappes est relativement élevé du fait que, une fois qu'un voisinage a été sélectionné pour faire partie de l'échantillon, il faut passer de nombreuses heures à obtenir la liste des ménages et des personnes d'un bureau ou service central et à s'assurer le soutien des leaders des voisinages. Le coût C_2 de 2 heures-personne associé aux individus comprend surtout des activités d'interview et de rappel. Le coût sur le terrain C_0 associé à un plan de sondage par EGSDO qui vérifie les deux conditions est donc (d'après la relation (7)) de 27,278 heures-personne, par rapport à un coût de 17,737 heures-personne (obtenu par la relation (8)) pour un plan de sondage par EGSDF qui satisfait aux deux exigences. L'EGSDF représente donc une économie substantielle de 35% en coûts sur le terrain.

À partir de la relation (9), nous calculons que, dans l'ensemble des 3 groupes d'âge, l'estimation x_{old}^{rel} , obtenue à partir d'un EGSDO de 419 grappes, du nombre de personnes qui manifestent des signes d'un désordre cognitif aurait une variance de 70,844, alors que x_{old}^{rel} , l'estimation analogue obtenue à partir d'un EGSDF ayant le même nombre de grappes (419), aurait une variance de 122,744, c'est-à-dire qu'elle serait de 42% plus élevée que la variance de l'estimation par EGSDO. Cependant, un plan de sondage par EGSDO ayant le même coût de travail sur le terrain qu'un EGSDF de 419 grappes échantillonnées ne permettrait d'échantillonner que 208 grappes (relation (11)). La variance de x_{old}^{rel} , calculée à partir d'un plan de sondage par EGSDO de 208 grappes échantillonnées serait estimée à 141,733, c'est-à-dire qu'elle serait de 15% plus élevée que la variance de l'estimation analogue par EGSDF ayant le même coût de travail sur le terrain. De plus, le plan de sondage par EGSDO à 208 grappes échantillonnées ne vérifierait pas les deux conditions posées sur les estimations.

6. ANALYSE

La méthodologie EGSDF analysée ici et dans des publications antérieures a été élaborée à partir d'une situation dans laquelle l'échantillonnage par grappes semblait clairement nécessaire mais où tout sous-échantillonnage de grappes était nettement contre-indiqué. Dans le cas de l'enquête de Shanghai sur la maladie d'Alzheimer et la démence, qui fait l'objet de notre analyse, les deux objectifs principaux étaient d'identifier un certain nombre d'individus dans chacun des 3 domaines (en l'occurrence des groupes d'âge) et d'obtenir des estimations qui se rapportent à chaque domaine particulier et satisfassent certaines conditions de précision. En fonction des résultats présentés ci-dessus à propos de cette enquête particulière, il semble que cette méthode pourrait entraîner des économies importantes quant au coût des activités sur le terrain sans mettre en péril les objectifs de l'enquête.

On pourrait avoir des doutes quant à l'applicabilité générale de cette méthodologie. Elle serait surtout utile dans des situations pour lesquelles le sous-échantillonnage des grappes est soit irréalisable soit trop coûteux, et où il n'est pas nécessaire de faire de dépistage pour déterminer si les individus appartenant à un des domaines cibles. (Dans l'ESMAD, les leaders des voisinages échantillonnés ont fourni une liste de toutes les personnes habitant dans le voisinage, ainsi que des renseignements sur les dates de naissance.) Des situations de ce genre pourraient se produire par exemple dans des enquêtes pour lesquelles les données sont extraites de dossiers par un personnel suffisamment familier avec ces dossiers pour en tirer de l'information, mais considéré incapable de les échantillonner sans une supervision coûteuse. Dans des situations de ce genre, comme nous l'avons dit plus haut, l'EGSDF peut constituer une solution de remplacement acceptable.

REMERCIEMENTS

La recherche qui a mené à cet article a été soutenue en partie par la subvention numéro I RO1 AG10327-01 du National Institute on Aging.

Parmi les 446 voisinages échantillonnés, 149 ont été désignés comme étant de type 1; dans 136 de ces voisinages, il y avait au moins une personne appartenant au groupe d'âge cible (55 ans et plus). Puisque seules les grappes de type 1 ont comme répondants admissibles toutes les personnes appartenant à chacun des 3 groupes d'âges cibles, on peut les utiliser pour estimer tous les paramètres nécessaires à la comparaison de l'efficacité en fonction des coûts de l'EGSDE et de l'EGSDO. Dans l'analyse qui va suivre, nous utiliserons les données provenant de ces 136 grappes pour illustrer de façon numérique comment on peut comparer l'efficacité en fonction des coûts de l'EGSDE et de l'EGSDO en utilisant les données "pilotes" disponibles. Pour chaque domaine h de cet échantillon de 136 grappes, les estimations des paramètres pertinents sont les suivantes:

Age	N_h	V^{N_h}	X_h	V^{X_h}	$\sum_{k < h}^{O_{h \times k}}$
55-64	10.985	.485	.125	2.991	0.000
65-74	8.088	.513	.360	2.357	0.190
75 +	3.478	.643	.456	1.665	0.296.

$$A_1 = 1.645 \times 0.485/2 = 0.3989$$

et

$$m'_1 = \left[0.3989 + \left((0.3989)^2 + \frac{1.400}{10.985} \right)^{\frac{1}{2}} \right] = 136.78 \approx 137.$$

De même, on trouve $m'_2 \approx 185$ et $m'_3 \approx 419$.

Supposons que, pour chacun des trois groupes d'âge, nous voulons estimer, à un niveau de confiance de 80% et avec une précision de 30%, la proportion de personnes qui manifestent des signes de dysfonctionnement cognitif. Ce dysfonctionnement est défini par l'obtention d'un résultat inférieur à 18 au Mini Mental State Examination (MMSE; Mini-examen de l'état mental), un test de dépistage du dysfonctionnement cognitif. Les mêmes données nous fournissent les estimations suivantes des paramètres nécessaires à la détermination du nombre de grappes échantillonnées requises pour satisfaire à ces exigences de certitude et de précision. En utilisant la relation (2), avec $M = 4,066$, $\epsilon = 0.30$, and $z_{1-\alpha/2} = 1.28$ nous trouvons les valeurs suivantes de m''_h :

$$m''_1 = 157; \quad m''_2 = 99; \quad m''_3 = 50.$$

D'après la relation (3), le nombre m_h de grappes requises pour que les deux conditions soient vérifiées dans chaque domaine est donné par:

$$m_1 = \max(137, 157) = 157; \quad m_2 = \max(185, 99) = 185; \quad m_3 = \max(419, 50) = 419.$$

Donc, pour qu'un plan de sondage par EGSDO satisfasse aux deux exigences, le nombre de grappes échantillonnées doit être de 419. De la même façon, un plan de sondage par EGSDÉ comportant 157 grappes échantillonnées de type 1, 28 de type 2 et 234 de type 3 vérifierait les deux conditions.

L'analyse ci-dessus révèle qu'à un coût équivalent à celui d'un EGSD de m^H grappes, la variance x_{old}^* (lorsqu'on laisse de côté la correction pour population finie) ne sera supérieure à celle obtenue dans un EGSD de m^H grappes que par un facteur de m^H/m^w au maximum, où m^w est une moyenne pondérée des grappes m_h requises dans chaque domaine pour que les exigences particulières au domaine soient satisfaites. Les poids sont ici les N_h , c'est-à-dire le nombre moyen d'individus dans chaque domaine particulier. Il faut remarquer aussi que la taille effective de l'échantillon dans un EGSD de coût équivalent à un EGSD diminue lorsque C_2/C_1 augmentera, C_2/C_1 étant essentiellement le rapport entre le coût de cueillette de l'information chez les individus échantillonnés et le coût de préparation des grappes échantillonnées en vue de l'enquête. Cette diminution semble intuitivement logique.

Les questions analysées ci-dessus sont illustrées dans la prochaine section à partir de données provenant de l'enquête de Shanghai sur la maladie d'Alzheimer et la démence.

5. L'ENQUÊTE DE SHANGHAI SUR LA MALADIE D'ALZHEIMER ET LA DÉMENCE

L'ESMAD, dont la planification remonte à 1986, a comme objectifs principaux: 1) l'estimation de la prévalence d'un ensemble de troubles physiques ou mentaux, au nombre desquels la maladie d'Alzheimer et d'autres maladies causant la démence, chez des personnes appartenant à chacun des trois groupes d'âge étudiés (55 à 64 ans, 65 à 74 ans, et 75 ans et plus), dans le district Jing-An de Shanghai en Chine et 2) l'identification d'environ 1,400 personnes dans chacun de ces trois groupes d'âge en vue de déterminer par la suite l'incidence de ces affections. Jing-An est l'un des douze districts qui constituent la ville de Shanghai. Il a été choisi comme secteur-cible à cause de sa population de personnes âgées, relativement importante et assez stable, et à cause de sa proximité de l'Institut de la santé mentale de Shanghai, qui était responsable du travail sur le terrain. Les constatations de cette étude ont été analysées par Zhang et coll. (1990) et par Yu et coll. (1989). Les questions de méthodologie sont examinées dans Levy et coll. (1988 et 1989).

Dans cette enquête, les grappes sont des unités administratives appelées *voisines*, constituées de ménages habitant des emplacements contigus. Ces ménages ont une structure politique et sociale bien définie. La stratégie retenue était de faire participer à l'identification et au recrutement de personnes admissibles les leaders des voisines sélectionnées pour faire partie de l'échantillon. À l'époque où l'enquête a été planifiée, il y avait 4,066 voisines dans le district Jing-An. Cette population particulière de Chinois âgés ou vieillissants avait généralement un faible niveau d'instruction. Ces gens avaient connu au cours de leur vie de nombreuses périodes de bouleversements politiques et de répressions (ex.: les seigneurs de la guerre, l'invasion japonaise, la Révolution culturelle), pendant lesquelles se faisaient remarquer ou être sélectionnés, avait souvent des conséquences néfastes. C'est pourquoi les membres de l'équipe de recherche, et surtout les membres chinois locaux qui connaissaient bien la population cible, avaient très forte impression que toute tentative de sous-échantillonnage dans les groupes d'âge cibles faisant partie des voisines échantillonnées compromettrait les taux de réponse et la coopération en général.

Le membre de l'équipe de recherche responsable du plan de sondage (PS), obligé d'utiliser l'échantillonnage par grappes à un seul degré et soumis à un échancier très serré pour l'élaboration du plan de sondage, a proposé une méthode heuristique qui permettrait, avec un assez grand degré de certitude, d'identifier 1,400 individus dans chacun des trois groupes d'âge cibles. Le plan de sondage résultant était essentiellement un EGSD dont l'échantillon était de 446 voisines. Pour des renseignements plus détaillés sur le plan de sondage, le lecteur est prié de se reporter aux publications de l'ESMAD mentionnées ci-dessus. Soulignons que ce plan n'a été choisi que pour des raisons heuristiques, bien avant l'élaboration de la théorie sur laquelle il est fondé.

Dans nombre de cas, la comparaison la plus importante, à coût équivalent, entre les deux plans de sondage porte sur leur capacité d'estimer le niveau d'ensemble X d'un caractère donné \mathcal{X} . La variance d'un estimateur x'_{ord} fondé sur un EGSDO de m^H grappes (le nombre requis pour satisfaire aux exigences de chaque domaine) est donnée par:

$$\text{Var}(x'_{ord}) = \frac{M^2}{m^2} \left(\frac{M - 1}{M - m^H} \right) \left(\sum_{h=1}^H \left(\sigma_{hx}^2 + 2 \sum_{k>h} \sigma_{h k x} \right) \right). \quad (9)$$

Cette formule n'est pas la formule habituelle pour la variance (Levy et Lemeshow 1991, chapitre 9), mais une forme algébriquement équivalente qui peut se comparer directement à la variance de x'_{tel} , obtenue par un plan de sondage par EGSDÉ avec m^H grappes échantillonnées (équation (6)). La différence entre ces deux variances est donnée par:

$$\text{Var}(x'_{tel}) - \text{Var}(x'_{ord}) = \frac{M^3}{m^3} \left(\frac{m^H - m^H}{m^H - m^H} \right) \left(\sum_{h=1}^H \left(\sigma_{hx}^2 + 2 \sum_{k>h} \sigma_{h k x} \right) \right). \quad (10)$$

Cette différence est supérieure ou égale à zéro (0), ce qui n'est pas surprenant puisqu'un EGSDO de m^H grappes produira toujours un échantillon global dont la taille sera plus grande que celle produite par un EGSDÉ ayant le même nombre de grappes.

Bien qu'un EGSDO de m^H grappes produise un estimateur x'_{ord} dont la variance est plus faible que celle de l'estimateur x'_{tel} résultant d'un EGSDÉ ayant le même nombre m^H de grappes, cette variance plus faible est obtenue à un coût plus élevé. C'est pourquoi il est plus utile de comparer le x'_{tel} obtenu à partir d'un échantillon avec le x'_{ord} obtenu à partir d'un échantillon de m^* grappes, où m^* est le nombre de grappes que l'on peut échantillonner par EGSDO à un coût équivalent à celui d'un plan de sondage par EGSDÉ comportant m^H grappes échantillonnées. Des équations (7) et (8) on déduit que m^* est donné par:

$$m^* = m^H \left(\frac{1 + \frac{C_2}{C_1} \sum_{h=1}^H \gamma_h N_h}{1 + \frac{C_2}{C_1} \sum_{h=1}^H N_h} \right). \quad (11)$$

Il est à noter que:

$$1) \quad m^* \leq m^H.$$

$$2) \quad \text{Lorsque } C_2/C_1 \rightarrow \infty, \text{ alors } m^* \rightarrow m^w$$

$$\text{où } m^w = \frac{\sum_{h=1}^H m_h N_h}{\sum_{h=1}^H N_h}.$$

$$3) \quad \text{Lorsque } C_2 \setminus C_1 \rightarrow 0, \text{ alors } m^* \rightarrow m^H$$

et

$$4) \quad m^* \text{ décroît de façon monotone lorsque } C_2/C_1 \text{ augmente, ce qui implique que } m^w \leq m^* \leq m^H.$$

On peut alors obtenir les résultats suivants par analyse combinatoire.

1. L'estimation par EGSDÉ x'_{tel} d'un chiffre de population X est donnée par:

$$(4) \quad x'_{tel} = \sum_H x'_h,$$

où x'_h est donné par

$$x'_h = (M/m_h) \sum_{i \in S_h} X_{hi}.$$

2. La moyenne $E(x'_{tel})$ et la variance $\text{Var}(x'_{tel})$ de x'_{tel} sont données par

$$(5) \quad E(x'_{tel}) = X,$$

$$(6) \quad \text{Var}(x'_{tel}) = \sum_H \frac{M^2}{m_h} \left(\frac{M - m_h}{M - 1} \right) \left(\sigma_{hx}^2 + 2 \sum_{k < h} \sigma_{hkx} \right).$$

Ces relations se déduisent de façon directe par analyse combinatoire.

4. COMPARAISON DES CÔÛTS DE L'EGSDO ET DE L'EGSDÉ

Nous pouvons comparer les coûts de l'EGSDO et de l'EGSDÉ en considérant la fonction de coût simple suivante, associée à l'EGSDO:

$$(7) \quad C_0 = C_1 m_H + C_2 m_H (N_1 + N_2 + \dots + N_H) = m_H \left(C_1 + C_2 \sum_H N_h \right),$$

où C_0 est le coût probable, C_1 est l'élément de coût associé aux grappes (*ex*: déplacement vers la grappe et de la grappe, obtention de la liste des unités de recensement qu'elle contient, pré-paration du matériel nécessaire au travail sur le terrain dans la grappe, *etc.*) et C_2 est l'élément de coût associé aux unités de liste (surtout le déplacement d'une unité à l'autre et les interviews). Il faut aussi remarquer que l'expression $\sum_H N_h$ est le nombre moyen d'unités sondées par grappe. En outre, tout au long de la présente analyse, les unités de liste sont les individus eux-mêmes. De façon analogue, le coût prévu associé à l'échantillonnage étagé est donné par:

$$(8) \quad C_l = C_1 m_H + C_2 (m_1 N_1 + m_2 N_2 + \dots + m_H N_H) = m_H \left(C_1 + C_2 \sum_H \gamma_h N_h \right),$$

où $\gamma_h = m_h/m_H$ (qui est ≤ 1). Le coût associé à l'EGSDÉ est donc inférieur ou égal à celui associé à un EGSDO du même nombre de grappes, la différence étant égale à

$$C_2 m_H \sum_H (1 - \gamma_h) N_h.$$

où

X_{hj} = le niveau de la variable \mathcal{X} pour l'individu j dans le domaine h de la grappe i ($j = 1, \dots, N_{hi}; i = 1, \dots, M$),

$$X_{hi} = \sum_{j=1}^{N_{hi}} X_{hj},$$

$$\bar{X}_h = \sum_{i=1}^I X_{hi}/M,$$

$$\sigma^2_{hx} = \sum_{i=1}^I (X_{hi} - \bar{X}_h)^2/M,$$

et

$$V^2_{hx} = \sigma^2_{hx}/X^2_h.$$

Pour que les deux exigences ci-dessus soient vérifiées dans chaque domaine, il s'ensuit que nous devrions échantillonner m_h grappes où, pour $h = 1, \dots, H$,

$$(3) \qquad m_h = \max(m'_h, m''_h).$$

Sans perte de généralité, nous pouvons renommer les domaines dans l'ordre croissant des m_h requis (c'est-à-dire $m_1 \leq m_2 \leq \dots \leq m_H$).

Enfin, pour que les deux exigences soient vérifiées dans chacun des H sous-domaines lorsqu'on emploie un plan de sondage par grappes à un seul degré ordinaire, le nombre m de grappes requises aux fins d'échantillonnage est m^H . Notons encore une fois que, dans un échantillonnage par grappes à un seul degré ordinaire, on échantillonne tous les individus dans chaque grappe sélectionnée. Donc, bien que les exigences relatives à la taille de l'échantillon soient satisfaites de façon minimale dans le domaine H (celui où le plus grand nombre de grappes sont requises), ces exigences sont amplement satisfaites dans les autres domaines, soit $1, \dots, H - 1$. Le fait d'avoir plus d'individus qu'il n'est vraiment nécessaire dans des domaines autres que H peut mener à une enquête pour laquelle les coûts associés au travail sur le terrain sont excessifs.

Lorsqu'on veut éviter cette dépense inutile liée à l'échantillonnage par grappes à un degré ordinaire, on a généralement recours à un plan de sondage par grappes à deux degrés où les tractions de sondage du deuxième degré (c'est-à-dire du deuxième échantillonnage) sont différentes d'un domaine à l'autre. Cependant, dans une situation où il n'est pas pratiquement réalisable de sous-échantillonner les grappes, Levy et coll. 1989, ont proposé d'utiliser une méthode appelée *échantillonnage par grappes à un seul degré étagé* (EGSDE), dans laquelle la règle d'*admissibilité* (c'est-à-dire la règle qui détermine quels individus peuvent faire partie de l'échantillon) peut varier d'une grappe échantillonnée à une autre, et dans laquelle les domaines échantillonnés ne sont pas les mêmes d'une grappe échantillonnée à une autre. Cet article de Levy a démontré l'utilité d'un échantillonnage à un seul degré étagé pour les enquêtes dont un des objectifs principaux est d'identifier un certain nombre d'individus dans chacun des sous-domaines pour un suivi longitudinal ultérieur. Dans le présent article, nous donnerons des propriétés caractéristiques des estimations provenant de ce type de plan de sondage et les comparerons à celles des estimations provenant d'un échantillonnage par grappes à un seul degré ordinaire.

2. FORMULATION STATISTIQUE

Supposons qu'une population est constituée de N individus et divisée en H sous-domaines mutuellement exclusifs, contenant chacun N_h individus ($h = 1, \dots, H$). Supposons de plus que la population est regroupée en M grappes, qui constitueront les unités d'échantillonnage de l'enquête. Enfin, nous supposons que l'échantillonnage des grappes se fera selon un échantillonnage par grappes à un seul degré ordinaire (c'est-à-dire un échantillonnage aléatoire simple de grappes, suivi d'une sélection de tous les individus de chaque grappe sélectionnée). Si nous voulons identifier, à un niveau de confiance de $100 \times (1 - \alpha)\%$, au moins n_h^* individus dans un domaine particulier h , alors on doit choisir le nombre m_h^* de grappes, donné par la formule suivante (Levy et coll. 1989):

$$m_h^* = \left\lceil A_h + \left(A_h^2 + \frac{N_h}{n_h^*} \right)^{1/2} \right\rceil, \tag{1}$$

où

N_{hi} = le nombre d'individus dans le domaine h de la grappe i , ($i = 1, \dots, M$),

$$N_h = \sum_{i=1}^M N_{hi}/M,$$

$$V_{N_h} = \sigma_{N_h}^2/N_h,$$

$$\sigma_{N_h}^2 = \sum_{i=1}^M (N_{hi} - N_h)^2/(M - 1),$$

z_α = le percentile de rang 100α de la distribution normale

et

$$A_h = |z_\alpha| \times V_{N_h}^{1/2}.$$

Dans ce qui précède, on suppose que les N_{hi} sont distribués normalement dans les M grappes. De plus, le nombre n_h^* d'individus requis dans le domaine h dépend de considérations statistiques ayant trait à la composante longitudinale de l'enquête. Il pourrait, par exemple, être calculé à partir du taux d'occurrence espéré de l'évènement sous étude pendant la période de suivi et de la précision requise pour l'estimation de ce taux d'occurrence. Si on veut aussi estimer à un niveau de confiance de $100 \times (1 - \alpha)\%$, et avec une précision de $100 \times \epsilon\%$ par rapport à sa vraie valeur dans chaque domaine, la valeur totale ou le niveau moyen d'une certaine variable \mathcal{Y} , alors le nombre M_h^* de grappes qu'il faudrait échantillonner dans le domaine h est donné par:

$$m_h^* = \frac{z_{1-\alpha/2}^2 V_{N_h}^2 + (M - 1) \epsilon^2}{z_{1-\alpha/2}^2 M V_{N_h}^2} \tag{2}$$

Echantillonnage par grappes à un seul degré dans les enquêtes prévalence-incidence: certaines questions soulevées par l'enquête de Shanghai sur la maladie d'Alzheimer et la démence

ZHISEN XIA, PAUL S. LEVY, ELENA S.H. YU, ZHENGYU WANG,
et MINGYUAN ZHANG¹

RÉSUMÉ

Nous considérons dans cet article le scénario d'une enquête par sondage ayant les deux objectifs principaux suivants: 1) l'identification, pour des études de suivi ultérieures, de n^* -sujets dans chacun des H sous-domaines et 2) l'estimation, au moment où on en est dans le déroulement de l'enquête, du niveau d'un caractère quelconque dans chacun de ces sous-domaines. Pour cette enquête, le plan de sondage doit se limiter à un échantillonnage par grappes à un seul degré, ce qui constitue une contrainte supplémentaire. Levy et coll. 1989, ont proposé une variante de l'échantillonnage par grappes à un seul degré, appelée *échantillonnage par grappes à un seul degré étagé* (EGSDE), comme moyen économique d'identifier n^* -sujets dans chacun des sous-domaines. Dans cet article-ci, nous étudions les propriétés statistiques de l'EGSDE pour l'estimation transversale du niveau d'un caractère dans la population. En particulier, la fiabilité d'estimations obtenues, à un coût donné, à l'aide de l'EGSDE est comparée à celle des estimations obtenues au même coût à l'aide de l'échantillonnage par grappes à un seul degré ordinaire (EGDO). Nous avons été motivés par les problèmes rencontrés au cours de la conception statistique de l'enquête de Shanghai sur la maladie d'Alzheimer et la démence (ESMAD), une étude épidémiologique de la prévalence et de l'incidence de la maladie d'Alzheimer et de la démence.

MOTS CLÉS: Échantillonnage par grappes à un seul degré; estimation de la prévalence; échantillonnage par grappes à un seul degré étagé; maladie d'Alzheimer; démence.

1. HISTORIQUE ET INTRODUCTION

Un grand nombre d'études ont à la fois une composante transversale, dans laquelle on estime les niveaux de variables quantitatives ou les prévalences de variables dichotomiques à l'aide d'un sondage, et une composante longitudinale, dans laquelle on utilise le même sondage pour définir une cohorte d'individus qui seront suivis pendant une période déterminée pour tenir compte d'événements ultérieurs. Ce type d'étude est particulièrement courant dans le domaine de l'épidémiologie, où on a besoin d'estimer la prévalence d'une maladie ou d'un état aussi bien pour la population étudiée dans son ensemble que pour des sous-groupes déterminés de cette population, et où on doit identifier, dans chacun des sous-groupes déterminés, un nombre suffisant d'individus qui n'étaient pas atteints au départ afin de permettre une estimation subséquente de l'incidence de l'état ou de la maladie (Kannel 1966).

La conception d'un plan de sondage à la fois efficace et économique constitue un défi puisque l'il faut sélectionner des nombres suffisants d'individus dans chaque domaine, souvent à l'aide d'une forme quelconque d'échantillonnage par grappes, de façon à assurer une estimation fiable de la prévalence et des incidences mentionnées ci-dessus. Dans le présent article, qui fait suite à une étude récente menée en Chine, nous examinons ces questions liées à la conception du plan d'échantillonnage lorsqu'on utilise un type particulier d'échantillonnage par grappes: l'échantillonnage par grappes à un seul degré.

¹ Zhisen Xia, Paul S. Levy et Zhengyu Wang, School of Public Health, University of Illinois at Chicago, Chicago, C.P. 6998, Chicago IL, 60680, E.-U., Elena S.H. Yu, School of Public Health, San Diego State University, San Diego, Californie, E.-U., Mingyuan Zhang, Institut de la santé mentale de Shanghai, Shanghai, République populaire de Chine.

- MORRISON, P. (1982). Alternatives for monitoring local demographic change. Dans *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee et H. Goldsmith). Beverly Hills, California: Sage, 97-111.
- MUNSELL, K. (1988). Towns cope with hazardous wastes. *Small Town*, 18, 30.
- PITTINGER, D. (1976). *Projecting State and Local Populations*. Cambridge, Massachusetts: Ballinger Press.
- POGGIE, J. (1972). Toward quality control in key informant data. *Human Organization*, 31, 23-30.
- RIVES, N. (1982). Assessment of a survey approach. Dans *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee et H. Goldsmith). Beverly Hills, California: Sage, 79-96.
- RIVES, N., et SEROW, W. (1984). *Introduction to Applied Demography Data Sources and Estimation Techniques*. Beverly Hills, California: Sage.
- SCHLEIFER, S. (1986). Trends in attitudes toward and participation in survey research. *Public Opinion Quarterly*, 50, 17-26.
- SEIDLER, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organizational analysis. *American Sociological Review*, 39, 816-831.
- SMITH, S. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, 81, 287-296.
- SMITH, S., et LEWIS, B. (1983). Some new techniques for applying the housing unit method of local population estimation: Further evidence. *Demography*, 20, 407-413.
- SMITH, S., et LEWIS, B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography*, 17, 323-339.
- SMITH, S., et MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 79, 282-289.
- STARSHINIC, D., et ZITTER, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. *Demography*, 5, 475-484.
- SUMMERS, G. (1982). *Industrialization. In Rural Society in the U.S.: Issues for the 1980s*, (Eds. D. Dillman and D. Hobbs). Boulder, Colorado: Westview Press, 164-174.
- SWANSON, D. (1989). Intervalles de confiance pour les estimations postcensitaires de la population: une étude de cas pour les petites régions. *Techniques d'enquête*, 15, 281-290.
- SWANSON, D., BAKER, B., et VAN PATTEN, J. (1983). Conceptual and practical features of the housing unit method. Presented at the Annual Meeting of the Population Association of America.
- U.S. BUREAU OF THE CENSUS (1978). State and local agencies preparing population estimates and projections: Survey of 1975-76. *Current Population Reports* P-25, 725. Washington, DC: U.S. Bureau of the Census.
- U.S. DEPARTMENT OF ENERGY (1988). Section 175 report. *Secretary of Energy's Report to the Congress Pursuant to Section 175 of the Nuclear Waste Policy Act, As Amended*. Washington, DC: Office of Civilian Radioactive Waste Management, U.S. Department of Energy.

Dans le cas de la procédure des experts locaux, cela signifie que l'on peut obtenir le nombre de ménages à partir des données des sociétés de service public et que ces dernières données peuvent servir de base de sondage. Une fois qu'un échantillon est choisi, l'efficacité de la procédure dépend du recrutement et des connaissances des experts locaux. Si ces critères peuvent être satisfaits, il semble que la procédure puisse être appliquée. L'étape suivante consiste à en déterminer la précision dans une application donnée.

Nous n'avons pu évaluer la précision de l'âge et d'autres données sur la composition estimées à l'aide de la procédure parce que ces données comme elles ont été recueillies dans le cadre du recensement décennal de 1990 ne sont pas encore disponibles au moment où ces lignes sont écrites. Toutefois, nous sommes encouragés par les résultats du test pour la population totale, qui montrent que la méthode peut donner des estimations très précises, même dans de petites régions rurales qui connaissent une évolution rapide.

REMERCIEMENTS

Les auteurs désirent remercier Mark Flotow et des critiques anonymes de versions antérieures du présent article pour leurs commentaires.

BIBLIOGRAPHIE

BROWN, R., GEERTSEN, H., et KRANNICH, R. (1989). Community satisfaction and social integration in a boomtown: A longitudinal analysis. *Rural Sociology*, 54, 568-586.

BYERLY, E. (1990). State and local agencies preparing population and housing estimates. *Current Population Reports Series P-25*, 1063. Washington, DC: U.S. Bureau of the Census.

CARLSON, J., SWANSON, D., et WILLIAMS, C. (1990). The development of small area socioeconomic data to be utilized for impact analysis: Rural, Southern Nevada. In *High Level Radioactive Waste Management Proceedings of the 1990 International Conference*. LaGrange Park, Illinois: American Nuclear Society, 985-990.

COCHRAN, W. (1977). *Sampling Techniques*, (3ième Edition). New York: Wiley.

ESPENSHADE, T., et TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.

FREUDENBURG, W. (1982). Social impact assessment. Dans *Rural Society in the U.S.: Issues for the 1980s*, (Eds. D. Dillman et D. Hobbs). Boulder, Colorado: Westview Press, 296-303.

KRANNICH, R., BERRY, E., et GREIDER, T. (1989). Fear of crime in rapidly changing rural communities: A longitudinal analysis. *Rural Sociology*, 54, 195-212.

KRANNICH, R., et GREIDER, T. (1984). Personal well-being in rapid growth and stable communities: Multiple indicators and contrasting results. *Rural Sociology*, 49, 541-552.

LEE, E., et GOLDSMITH, H. (1982). Evaluation and synopses. Dans *Population Estimates: Methods for Small Area Analysis*, (Eds. E. Lee and H. Goldsmith). Beverly Hills, California: Sage, 113-118.

LOWE, T., PITTINGER, D., et WALKER, J. (1977). Making the housing unit method work: A progress report. Présenté à l'Annual Meeting of the Population Association of America.

LOWE, T., WEISSER, L., et MYERS, B. (1984). A special consideration in improving housing unit method estimates: The interaction effect. Présenté à l'Annual Meeting of the Population Association of America.

MARTIN, J., et SEROW, W. (1979). Virginia's state-local cooperative program: conflict and cooperation in producing population estimates. State Government, 182-186.

population du recensement de 1990 qui figure au tableau 1 pour Pahrump est celle qui est donnée pour cette division du comté de Nye.

Il existe d'autres différences entre les estimations et les chiffres du recensement de 1990. La date officielle des chiffres du recensement est le 1^{er} avril; les estimations sont faites pour le 15 juillet. En fonction de cette différence, on croit que les effets saisonniers sont très faibles pour les collectivités étudiées. À l'exception du départ de certaines personnes qui se dirigent vers des lieux plus chauds et qui peuvent avoir été dénombrées dans la zone d'étude parce qu'elles n'avaient pas de lieu habituel de résidence ailleurs, il n'y avait pas de flots connus de migration d'une certaine importance entre avril et juillet. De même, les autres composantes du mouvement de la population étaient peu importantes.

Si le Bureau avait trouvé des itinérants sans domicile habituel ailleurs, il est probable que la procédure d'estimation n'aurait pas permis de les dénombrer. Ces différences auraient aussi une incidence sur le nombre de logements. Si un itinérant, reconnu comme un habitant de la collectivité aux fins du recensement décennal était trouvé dans un véhicule de plaisance, ce dernier serait inclus par le Bureau dans le parc de logements "autre" de la collectivité. De tels logements ne seraient pas inclus dans les données provenant des dossiers sur les compteurs utilisés pour le service résidentiel d'électricité.

Nous croyons, cependant, que de tels cas sont rares et, de plus, la comparaison d'une population qui réside dans des logements avec une population qui réside surtout dans des logements mais aussi, dans une certaine mesure, dans des logements de groupe n'entraîne pas de confusion dans les résultats du test.

Les résultats du test de précision sont aussi résumés au tableau 1, avec les estimations "faible" et "élevée" correspondant à l'intervalle de confiance à 95 pour cent établi autour de la population estimée de chacune des collectivités. La population estimée est très proche de la population déclarée par le Bureau. Dans l'ensemble, la moyenne de la valeur absolue des différences est de 86 personnes et la moyenne de la valeur absolue des différences en pourcentage est de 1,7%.

Les trois intervalles de confiance comprenant la population du recensement de 1990 dans chacune des trois collectivités. Cette conclusion présente un intérêt particulier parce que les intervalles sont relativement peu étendus pour un niveau de confiance à 95 pour cent. En moyenne, l'étendue, telle que mesurée entre la population estimée et l'une ou l'autre des deux limites, correspond à 7,2 pour cent de la population estimée. Cela laisse supposer que les intervalles de confiance construits autour des estimations obtenues à l'aide de cette variante de la MUL sont significatifs, même en présence d'un niveau inconnu d'erreur non due à l'échantillonnage.

La population de deux des trois collectivités est sous-estimée. Dans le cas de Pahrump, la technique d'estimation n'a apparemment pas été en mesure de tenir compte de toute la croissance récente qui semble déborder de la vallée de Las Vegas dans la région de Pahrump. On ne sait pas dans quelle mesure cette erreur était due à des logements manquants dans la base de sondage et quelle proportion de l'erreur était imputable à une sous-estimation de la valeur du PPH pour Pahrump.

10. RÉSUMÉ

Bien que la procédure des experts locaux puisse ne pas fournir des estimations satisfaisantes de la population dans toutes les petites régions rurales (p. ex. les lieux de villégiature où l'on compte une proportion élevée d'unités de logement saisonnières et d'unités louées par des particuliers) elle semble prometteuse si l'on se base sur les données pour la région incluse dans la présente étude. Comme dans le cas de toute technique d'estimation, le critère-clé pour déterminer si la technique pourrait être appliquée ailleurs dépend de la possibilité d'obtenir les données nécessaires et de mettre la procédure en oeuvre compte tenu des fonds disponibles.

Posons

P = population estimée dans les ménages,
 N = nombre de ménages,
 PpH = nombre estimé de personnes par ménage.

Alors

$$\begin{aligned} \text{limite inférieure } (P) &= (N) * (PpH - (t^{n-2}, \alpha/2) * (se)), \\ \text{limite supérieure } (P) &= (N) * (PpH + (t^{n-2}, \alpha/2) * (se)), \end{aligned}$$

où

n = nombre de ménages échantillonnés,
 α = niveau de signification désiré,
 se = erreur type du PpH estimé,

$$t^{n-2} = (\alpha/2) \text{ 100e percentile de la distribution } t, \text{ avec } (n - 2) \text{ degrés de liberté.}$$

Comme exemple, si l'on utilise un niveau de signification de .05, l'intervalle de confiance à 95% correspondant pour la population estimée de Pahrump en 1990 (7,190) est

$$\begin{aligned} \text{limite inférieure} &= 6,810 = (3,224) * (2.23 - (1.96 * .06)), \\ \text{limite supérieure} &= 7,569 = (3,224) * (2.23 + (1.96 * .06)). \end{aligned}$$

9. TEST DE PRÉCISION

Avant de passer aux résultats du test, qui figurent aussi au tableau 1, nous devons traiter de certaines restrictions qui s'appliquent aux données.

La question la plus problématique pour ce qui est de comparer les estimations avec les résultats du recensement de 1990 est due au fait que le Bureau of the Census ne reconnaît pas les "districts fiscaux" comme des limites administratives pour les collectivités dans la zone d'étude. Cela signifie qu'on doit utiliser la géographie "statistique" du Bureau, ce qui exige certains ajustements pour que la géographie utilisée aux fins de l'analyse des incidences sur l'environnement corresponde à celle utilisée par le Bureau. En fonction de ces ajustements, on sait que la région définie comme Amarogosa Valley aux fins de l'analyse de l'incidence sur l'environnement ne correspond pas à la Amarogosa Valley Census Division du comté de Nye utilisée par le Bureau en ce sens que la définition de l'étude comprend la Crystal Census Division du comté de Nye. Heureusement, il s'agit d'un cas où l'on peut combiner deux éléments de géographie statistique utilisés par le Bureau afin qu'ils correspondent presque exactement à l'élément géographique employé dans l'analyse des incidences sur l'environnement. Ainsi, les chiffres de population du recensement de 1990 qui figurent au tableau 1 pour Amarogosa Valley comprennent la Crystal Division ainsi que la Amarogosa Valley Division. "Beatty" est une autre région que l'on sait différer pour ce qui est des éléments géographiques. Elle est définie à la fois comme une Census Designated Place (CDP) et comme la Beatty Census Division du comté de Nye par le Bureau. Dans ce cas, c'est la Census Designated Place qui correspond de très près à la définition de Beatty utilisée dans l'étude des incidences sur l'environnement. L'effectif, selon le recensement de 1990, pour Beatty qui figure au tableau 1 est donc le chiffre qui correspond à la CDP de Beatty.

La troisième collectivité, Pahrump, est désignée comme une Census Division du comté de Nye. Cet élément de géographie statistique utilisé par le Bureau est virtuellement identique à celui qui est employé dans l'étude des incidences sur l'environnement. Par conséquent, la

7. RÉSULTATS

Le premier produit de données est le nombre de ménages, qui est obtenu directement à partir des dossiers actifs sur les compteurs, triés et classés par les employés des sociétés de service public. Le tableau 1 montre ces chiffres classés par collectivité ainsi que d'autres résultats dont on traitera plus loin.

Le tableau fournit aussi le *PPH* estimé, qui est tiré du nombre global de personnes relevées, dans les unités de logement occupées de l'échantillon, par les experts locaux. On trouve aussi dans ce tableau la population estimée dans les ménages de chaque collectivité; ce chiffre a été obtenu en appliquant la formule de la *MUL* aux composantes ménage et *PPH*. On n'a pas relevé de logement de groupe dans l'une quelconque des collectivités étudiées.

8. MESURE DE L'IMPRÉCISION DANS LES ESTIMATIONS

Le fait qu'on puisse produire des intervalles de confiance constitue un avantage important des estimations basées sur l'échantillonnage aléatoire. Rives (1982) préconise cette méthode. Toutefois, il n'a pas considéré la possibilité d'avoir recours à des experts locaux et croyait que sa suggestion ne serait suivie que dans des circonstances exceptionnelles à cause du coût élevé des enquêtes traditionnelles. Ce fait a aussi été remarqué par Morrison (1982) et Lee et Goldsmith (1982) dans leur étude critique de la suggestion de Rives.

Dans le cas de la procédure des experts locaux, la "statistique" est la valeur du *PPH* qui, en pratique, varierait d'un échantillon à l'autre. Toutefois, nous sommes moins intéressés aux valeurs du *PPH* qu'à l'estimation de la population, nous utilisons donc une transformation simple introduite par Espenshade et Tayman (1982) et utilisée plus récemment par Swanson (1989) pour placer les intervalles de confiance produits, à l'origine, pour la valeur du *PPH* pour une collectivité donnée autour de chacune des estimations de la population de la collectivité.

Tableau 1
Caractéristiques de l'échantillon et résultats du test de précision*

Collectivité	Ménages	Estimation pour 1990			Chiffre du recensement de 1990	Intervalle de confiance à 95%	
		<i>PPH</i>	<i>SE</i>	Population		Inférieur	Supérieur
Amargosa Valley	326	2.58	.11	841	853	771	911
Beatty	672	2.43	.10	1,633	1,623	1,501	1,765
Pahrump	3,224	2.23	.06	7,190	7,425	6,810	7,569

* Les données estimées et les intervalles de confiance sont produits à l'aide des procédures décrites dans le texte. Les chiffres du recensement de 1990 sont tirés du tableau 3 de "1990 Census Extract, Nevada, Public Law 94-171 Data," du 11 février 1991, document distribué par Betty McNeal, Nevada State Data Center Librarian, Nevada State Library and Archives, Capitol Complex, Carson City, Nevada 89710. Le chiffre pour la région "Amargosa Valley" est constitué de la population déclarée en 1990 pour l'Amargosa Valley Division (761) et pour la Crystal Valley" est constitué de la population déclarée en 1990 pour l'Amargosa Valley Division (761) et pour la Crystal Valley" est tiré des données pour la Beatty Census Division (92) du comté de Nye. Le chiffre pour la région "Beatty" est tiré des données pour la Beatty Census Designated Place et le chiffre pour la région "Pahrump" est tiré des données pour la Pahrump Division du comté de Nye.

comprenaient les genres et les emplacements des commerces et des zones résidentielles. Quatre types de logements distincts ont été définis à l'aide des lignes directrices élaborées par le U.S. Bureau of the Census.

Après l'étude préliminaire, le réseau routier ainsi que d'autres caractéristiques ont été tracés sur une carte pour chaque collectivité. À l'aide de ces cartes et des dossiers des sociétés de service public, des représentants des sociétés d'électricité desservant le sud du comté de Nye ont déterminé l'emplacement et le genre de logement, s'il y en a, associées à tous les raccordements électriques actuels. Ces renseignements ont été ajoutés au dossier sur les unités de logement constitué à partir des dossiers des sociétés de service public pour chaque collectivité. À cause de l'absence de dossiers adéquats pour les services publics dans le cas de Indian Springs, les renseignements sur les logements pour cet endroit ont été recueillis au moyen d'une enquête superficielle, un examen systématique, lot par lot, des unités de logement effectué par des équipes travaillant en automobile (Lowe, Pitenger et Walker 1977). C'est pourquoi on ne tient pas compte d'Indian Springs dans les résultats de l'essai mentionnés dans le présent article. Le travail préliminaire sur le terrain a montré que l'on pouvait s'attendre à des différences importantes dans le *PPH* entre les collectivités dans la zone d'étude. On a donc tiré séparément, pour chaque collectivité, d'après le nombre d'unités de logement qu'on y retrouve, un échantillon aléatoire d'unités provenant du dossier sur les unités de logement. Nous avons utilisé une méthode prudente pour déterminer la taille de l'échantillon dans chaque collectivité. Cette méthode suppose une marge d'erreur de 5%, un niveau de signification de .05 et un intérêt dans une variable dichotomique avec une distribution 50-50 (Cochran 1977). Une fois la taille initiale déterminée, nous avons ajouté un autre 15% pour tenir compte des cas manquants. La taille finale des échantillons pour Amargosa Valley était de 175 unités de logement, pour

Beatty de 222 et pour Pahrump de 355.

Au début, les experts locaux ont été trouvés par l'intermédiaire du réseau de personnes avec qui on entretenait des rapports, en fonction de leur expérience dans des manifestations d'initiales locales et de la connaissance qu'ils avaient des habitants des collectivités. Chaque personne qui pouvait éventuellement être embauchée à titre d'expert a été interviewée et on lui a demandé de remplir un formulaire conçu afin d'évaluer sa compétence. Une explication écrite du projet et des instructions précises sur le processus de collecte des données ont été fournies et on en a discuté. Les personnes choisies comme experts locaux ont reçu des instructions au sujet de la confidentialité. Pour ce projet, nous avons trouvé que les agents relèvent de compteurs employés par les sociétés de service public constituaient une bonne source d'experts locaux. On a fourni à ces derniers l'échantillon d'unités de logement pour la collectivité dans laquelle ils devaient travailler. Dans la majorité des cas, deux experts locaux travaillaient ensemble, ce qui permettait de vérifier la précision des informations au moment de leur inscription. Pour chaque unité, les experts locaux ne communiquaient au chercheur que le nombre de personnes dans le ménage au 15 juillet 1990, l'âge (à l'aide de huit groupes d'âge) et le sexe de chaque membre du ménage ainsi que le statut, par rapport à la retraite, de tout membre du ménage âgé de cinquante ans ou plus. Si l'un ou l'autre des deux experts locaux n'était pas certain de la composition d'un ménage donné, on communiquait avec un autre membre de la collectivité pour confirmer les données. Dans le cas où la composition d'une partie quelconque du ménage ne pouvait être confirmée, on inscrivait "données inconnues" pour toute l'unité. Les données étaient inscrites sur un formulaire sur lequel figuraient les unités échantillonnées désignées à l'aide d'un numéro d'attribut (établi en fonction de l'emplacement de l'unité sur la carte des unités de logement). Le numéro attribué au compteur d'électricité de l'unité ainsi que le genre d'unité de logement. On considère que toutes les unités résidentielles, y compris celles désignées comme "brûlées" ou détruites d'une autre façon, inoccupées ou "enlevées de l'emplacement pour maisons mobiles ou pour roulottes" (dans le cas des maisons mobiles et des roulottes) font partie de l'échantillon final. Les unités désignées comme "ne constitue pas une résidence" ont été éliminées de la base de sondage et ne figuraient pas dans l'échantillon. Il y avait quelques unités à propos desquelles on ne connaissait rien. Ces unités ne sont pas incluses dans l'échantillon final, ce qui peut causer un certain biais.

Un autre avantage qu'offre l'emploi des données des entreprises de service public est que les données utilisées pour obtenir le nombre total de ménages peuvent aussi être employées comme base de sondage complète à partir de laquelle on peut tirer des échantillons afin d'obtenir une estimation du nombre moyen de personnes par ménage (*PPH*). La collecte traditionnelle de données pour obtenir ce genre de renseignements à l'aide d'un échantillon prend l'une des trois formes suivantes: interview postale, interview téléphonique et interview sur place. Nous proposons d'utiliser à la place de ces méthodes des "experts locaux" afin de minimiser à la fois le coût et le dérangement.

4. EXPERTS LOCAUX

La procédure des experts locaux (aussi connue sous le nom de procédure des informateurs-clés) utilisée pour obtenir des renseignements à propos d'une collectivité est bien connue dans le domaine de l'anthropologie culturelle. On reconnaît généralement qu'il s'agit de "se fier à un petit nombre de participants bien renseignés qui observent et expriment clairement les rapports sociaux pour le chercheur (TRADUCTION)" (Seidler 1974: 816). De plus, Fogie (1972) trouve que lorsque les questions posées sur le terrain se rapportent à des phénomènes publics concrets, directement observables et non sujets à controverse, les experts locaux sont une source très fiable et précise de renseignements.

Il y a deux problèmes-clés quand on utilise la procédure des experts locaux conjointement avec les dossiers des sociétés de service public et la MUL. Le premier consiste à trouver et à recruter des personnes qui sont vraiment des experts locaux sur la composition des ménages qui font partie de l'échantillon. Le deuxième est la question de pouvoir obtenir des renseignements, permettant de repérer les ménages, qui sont familiers aux experts locaux (p. ex., une adresse de voirie et le nom du membre responsable du ménage plutôt que le code de facturation d'une société de service public).

5. ETUDE DE CAS

L'activité de collecte de données sur laquelle nos estimations de la population sont basées fait partie d'un programme visant à évaluer les caractéristiques socio-économiques de collectivités situées près de Yucca Mountain, Nevada, site proposé pour un dépôt de déchets nucléaires situé dans une formation géologique (U.S. Department of Energy 1988). Les données feront partie de l'ensemble de renseignements utilisés dans une analyse complète des incidences sur l'environnement du dépôt proposé. Yucca Mountain se situe dans le comté de Nye, environ 90 miles au nord-ouest de Las Vegas dans une région désertique peu peuplée. L'analyse des incidences sur l'environnement se concentre sur les collectivités qui se trouvent dans un rayon de cinquante miles du site de Yucca Mountain. Les zones d'étude comprennent les collectivités non constituées de Amargosa Valley, Beatty et Pahrump dans le sud du comté de Nye et de Indian Springs dans le comté de Clark. Nous utilisons les limites établies à des fins fiscales par les comités de comté pour préciser les limites des comtés aux fins de l'analyse des incidences sur l'environnement.

6. DONNÉES ET MÉTHODES

Au cours d'une phase préliminaire de la recherche, nous sommes entrés en communication avec des dirigeants et des habitants des collectivités. Ces rapports nous ont permis d'établir un réseau qui a facilité, par la suite, la collecte de données. Des notes décrivant la structure générale de chaque localité dans la zone d'étude ont été prises sur le terrain. Ces notes

région rurale. Toutefois, des restrictions relatives aux données ainsi qu'un désir de précision limitent considérablement la gamme de techniques qui peuvent être utilisées et, d'une façon réaliste, font ressortir une seule technique: la MUL (Smith 1986; Smith et Mandell 1984; Lowe, Weisser et Myers 1984; Swanson, Van Patten et Baker 1983; Smith et Lewis 1983, 1980).

3. LA METHODE DES UNITES DE LOGEMENT

Le concept de la MUL repose sur le fait que presque tout le monde dort sous un certain genre d'abri. Le U.S. Bureau of the Census, par exemple, choisit de définir deux classes d'abris: les logements de groupe et les unités de logement. On classe toutes les personnes dans l'une ou l'autre de ces catégories d'abris. Dans la MUL, on considère que ces abris peuvent être déterminés, comptés et classés comme occupés ou vacants. De plus, tous les abris occupés doivent avoir un nombre déterminé d'occupants. Par conséquent, la population de tout endroit donné doit être égale à la somme du produit du nombre d'unités de logement multiplié par le taux d'occupation multiplié par le nombre moyen de personnes par unité de logement occupée (ménages) plus le nombre de personnes dans les logements de groupe. Les quatre éléments de la MUL donnent une identité démographique exacte, qui permet d'obtenir la population d'un endroit particulier de la façon suivante:

$$P = [(H) * (O) * (PPH)] + GQ,$$

où

P = population totale,
 H = nombre total d'unités de logement,
 O = proportion d'unités de logement occupées,
 PPH = nombre moyen de personnes par ménage,
 GQ = population des logements de groupe.

Quand on utilise la MUL, la question-clé en matière de précision est de déterminer chacune des composantes. De plus, comme Smith (1986: 245-246) le fait remarquer:

«La méthode des unités de logement est une forme robuste, complète et extrêmement flexible d'estimation de la population avec un certain nombre de caractéristiques qui la rendent utile pour l'analyse de petites régions. Elle n'est pas limitée à une seule technique ou à un seul genre de données; elle peut, plutôt, utiliser un certain nombre de techniques et de sources de données différentes, y compris celles qui peuvent être applicables dans une région mais pas dans une autre. (TRADUCTION)»

Comme Smith (1986) l'a aussi fait remarquer, on utilise deux méthodes principales pour produire l'élément «nombre de ménages» de la MUL. Une de ces méthodes est basée sur des mesures de l'activité de construction et sur l'estimation d'un taux d'occupation; l'autre utilise des données des sociétés de service public, comme celles sur les clients du service résidentiel d'électricité. Un avantage important de la deuxième méthode est qu'elle peut fournir directement le nombre de ménages, ce qui élimine ou réduit de façon marquée un certain nombre d'inexactitudes possibles dans les données, y compris le besoin d'estimer les décalages temporels entre le moment où les permis sont émis et celui où les unités de logement sont terminées, les taux d'achèvement, les démolitions, les conversions et les taux d'occupation. Starinic et Zitler (1968) ainsi que Rives et Serow (1984) trouvent que la méthode des «données des entreprises de service public» utilisée pour la MUL est avantageuse, bien qu'ils admettent aussi qu'elle ait certaines limitations.

particulièrement adaptées pour de petites régions rurales, qui exploitent pleinement les données disponibles, qui sont moins coûteuses et, dans de nombreux cas, qui dérangent moins que les enquêtes aéroportées, les enquêtes téléphoniques et les enquêtes postales. Nous croyons que la variante de la MUL que nous proposons dans le présent article contribue à ce genre de progrès méthodologique.

La variante de la MUL que nous décrivons dans le présent article combine deux méthodes qui sont, en elles-mêmes, bien connues. Cependant, elles ont surtout été élaborées isolément l'une de l'autre, tout comme de la MUL. Ce sont: (1) l'échantillonnage aléatoire et (2) les interviews auprès d'"experts locaux". Comme nous en discutons plus loin, ces méthodes combinées avec la MUL peuvent nous mener à un moyen d'obtenir la taille de la population et, éventuellement, des données sur la composition nécessaires pour répondre aux besoins en information des projets d'évaluation environnementale et d'autres activités touchant les petites régions rurales.

2. CONSIDÉRATIONS LORS DE L'ÉVALUATION DES IMPACTS DANS DE PETITES RÉGIONS RURALES

L'implantation de nouvelles usines ou de nouvelles industries dans des régions rurales exige généralement une main-d'œuvre plus nombreuse que celle qui est disponible dans la région locale. On peut s'attendre à ce que la croissance de la population dans les collectivités situées près du site varie selon la taille du projet et le nombre d'employés qui seront embauchés pour construire l'installation puis pour l'exploiter et en assurer l'entretien une fois qu'elle sera terminée. Que l'on prévise ou non des augmentations rapides dans le nombre global de personnes, ou qu'il y ait ou non des changements importants dans la distribution par âge et par sexe, la structure modifiée de la population aura un effet sur le genre et sur la quantité de services publics nécessaires (Summers 1982). Ainsi, pour effectuer des évaluations environnementales, il faut disposer de renseignements sur les augmentations prévues dans les inscriptions dans les écoles, dans les besoins en logements, dans les besoins de services de santé et dans d'autres services. Toutefois, avant que l'on puisse faire de telles projections, on doit déterminer des renseignements sur la population actuelle de la région touchée afin de disposer d'une population "de départ" à des fins de prévision (Carlson, Williams, et Swanson 1990; Pittenger 1976; U.S. Department of Energy 1988).

La compréhension des principaux facteurs ayant un effet sur la distribution des personnes dans des régions rurales isolées est critique pour établir des profils et des projections démographiques. Il est probable que ces collectivités ont déjà connu des périodes de forte croissance et de déclin (Kranich et Greider 1984). Il se peut que l'évolution historique du mouvement de la population, ainsi que les tendances actuelles, diffèrent considérablement des moyennes obtenues à partir des renseignements sur le comté dans son ensemble ou même à partir d'autres régions incluses dans le comté. Cette situation présente un problème spécial parce qu'habituellement des renseignements démographiques précis ne sont disponibles que pour les années y compris les recensements fédéral est effectué. Toutefois, les données du recensement, au cours desquelles un recensement fédéral est effectué. Puisque le coût est habituellement localités non constituées comptant une faible population. La possibilité de réaliser des recensements spéciaux ou de grosses enquêtes-échantillons, particulièrement sur une base régulière, est souvent exclue, même dans de petites régions rurales. Un problème additionnel associé à de tels dénominations est dû au fait que pour les réaliser il faut employer des intervieweurs qui doivent communiquer avec chaque ménage, ce qui prend du temps et nuit à la vie privée des répondants et ajoute au fardeau de dérangement qui peut déjà être élevé pour les habitants de ces localités (Brown, Geertsen et Kranich 1989; Kranich, Berry et Greider 1989; Schleifer 1986). En principe, plusieurs techniques pourraient être employées pour estimer la taille de la population actuelle d'une petite

Une variante de la méthode des unités de logement pour estimer la population de petites régions rurales: une étude de cas portant sur la procédure des experts locaux

LINDA K. ROE, JOHN F. CARLSON et DAVID A. SWANSON¹

RÉSUMÉ

Dans le présent article, on examine la pertinence d'une procédure à base d'enquêtes pour estimer la population dans de petites régions rurales. La procédure est une variante de la méthode des unités de logement. Elle fait appel aux services d'experts locaux embauchés pour fournir des renseignements à propos des caractéristiques démographiques de ménages choisis aléatoirement dans des bases de sondage de quartiers d'habitation élaborées à partir des dossiers des sociétés de service public. La procédure ne dérange pas trop et elle est moins dispendieuse que les opérations traditionnelles de collecte de données au moyen d'enquêtes. Comme la procédure est basée sur l'échantillonnage aléatoire, on peut construire des intervalles de confiance autour de la population estimée au moyen de cette technique. On fournit les résultats d'une étude de cas portant sur l'estimation de la population totale dans trois collectivités non constituées situées dans une région rurale du sud du Nevada.

MOTS CLÉS: À base d'enquêtes; dossiers des sociétés de service public; intervalles de confiance Nevada.

1. INTRODUCTION

Dans sa dernière enquête sur les organismes d'Etat et locaux qui préparent des estimations de la population et des logements, le U.S. Bureau of the Census a trouvé qu'environ 89 pour cent des organismes interrogés utilisent la méthode des unités de logement (MUL) (Byerly 1990). Au cours d'une enquête antérieure on avait aussi constaté que cette méthode était largement utilisée (U.S. Bureau of the Census 1978). On a trouvé que la méthode fournissait des estimations précises de la population totale (Lowe, Pittenger et Walker 1977; Lowe, Weisser, et Myers 1984; Smith et Lewis 1980, 1983; Smith et Mandell 1984) ainsi qu'une base conceptuelle et pratique solide pour un système d'évaluation municipal (Martin et Serow 1979; Rives et Serow 1984; Swanson, Baker et Van Patten 1983). Une des caractéristiques importantes de la MUL est qu'elle peut être mise en oeuvre sous différentes formes, ce qui permet de l'adapter à une gamme d'environnements de données (Swanson, Baker et Van Patten 1983). Cette possibilité d'adapter la méthode a été exploitée surtout par des centres démographiques infra-nationaux à des fins de partage des recettes et pour des programmes connexes (Martin et Serow 1978; Swanson, Baker et Van Patten 1983). Toutefois, comme Rives (1982) le fait remarquer, la méthode pourrait être utilisée dans d'autres champs d'activité.

Considérons, comme exemple, le cas des énoncés des incidences environnementales. Les préoccupations à propos de questions juridiques et environnementales ont entraîné des décisions suite auxquelles des installations impopulaires ont été situées dans des régions rurales peu peuplées pour lesquelles on ne dispose généralement pas de données du recensement ou d'autres données socio-économiques (Freudenburg 1982; Brown, Geertsen et Kramlich 1989; Munsell 1988). Par conséquent, il est devenu nécessaire d'élaborer des méthodes d'enquêtes,

¹ Linda K. Roe et John F. Carlson, Science Applications International Corporation, 101 Convention Center Drive, Las Vegas, Nevada 89109; David A. Swanson, Center for Social Research and Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447-003, USA.

- WITTES, J. T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 93-97.
- Applique des systèmes multiples et l'indépendance pour estimer la taille d'une population d'enfants atteints d'une anomalie congénitale et d'autres problèmes.
- WITTES, J. T., COLTON, T., et SIDEL, V. W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.
- Applique des systèmes multiples et l'indépendance pour estimer la taille d'une population d'enfants atteints d'une anomalie congénitale.
- WITTES, J. T., et SIDEL, V. W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases*, 21, 287-301.
- Utilise la méthode de saisie-resaisie pour estimer le nombre de patients dans les hôpitaux qui utilisent la méthicilline.
- WOLTER, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- Décrit la version de base de la méthode dual telle qu'elle est utilisée pour le recensement de 1980, y compris l'élimination des enregistrements erronés.
- WOLTER, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.
- Utilise le rapport de masculinité pour obtenir un estimateur modifié de saisie-resaisie pour des données avec stratification selon le sexe, en supposant soit un risque relatif commun, soit l'indépendance dans une strate. Applique la méthode à des données sur les animaux et décrit l'application aux données du recensement.
- WOLTER, K. M. (1991). Policy Forum: Accounting for America's uncounted and miscounted. *Science*, 253, 12-15.
- Décrit les procédures de redressement des chiffres du recensement de 1990 et explique pourquoi elles sont défendables du point de vue statistique.
- WOLTER, K. M., et CAUSEY, B. D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.
- Examine, à l'aide de simulations, les améliorations apportées par le redressement synthétique par rapport aux chiffres du recensement pour des petites régions.
- ZALAVSKY, A. M., et WOLFGANG, G. S. (1990). Triple system modeling of census, post-enumeration survey, and administrative list data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.
- Applique divers modèles log-linéaires et connexes de la dépendance pour analyser les données de système triple provenant de la répétition générale du recensement qui a eu lieu à Saint Louis.

- SEBER, G.A.F. (1965). A note on the multiple-recapture census. *Biometrika*, 52, 249-259.
- Estimation à partir de données obtenues par resaisies multiples de populations ouvertes avec des paramètres liés au temps.
- SEBER, G.A.F. (1973). *The estimation of Animal Abundance and Related Parameters*. New York: Hafner. Deuxième édition (1982). New York: Macmillan.
- Renferme une revue à jour des techniques de saisie-resaisie et de leurs prolongements pour les populations d'animaux, avec l'accent mis sur les applications.
- SEBER, G.A.F. (1982). Capture-recapture methods. Dans *Encyclopedia of Statistical Sciences*, (Volume 1), (éds. S. Kotz et N.L. Johnson). New York: Wiley, 367-374.
- Examine la méthode de saisie-resaisie pour des populations tant fermées qu'ouvertes et fournit un guide pour des applications dans le domaine de la faune et des pêches.
- SHAPIRO, S. (1949). Estimating birth registration completeness. *Journal of the American Statistical Association*, 45, 261-264.
- Décrit l'utilisation de données du recensement des E.-U. de 1940 pour vérifier, à l'aide de l'estimateur de Chandrasekar-Deming, dans quelle mesure les enregistrements de la naissance sont complets.
- SHAPIRO, S. (1954). Recent testing of birth registration completeness in the United States. *Population Studies*, 8, 3-21.
- Décrit l'utilisation de données des recensements des E.-U. de 1940 et de 1950 pour vérifier, à l'aide de l'estimateur de Chandrasekar-Deming, dans quelle mesure les enregistrements de la naissance sont complets.
- SIRKEN, M.G. (1978). Dual systems estimators based on multiplicity surveys (avec discussion). Chapitre 4 dans *Developments in Dual System Estimation of Population Size and Growth*, (éd. K. J. Krótki). Edmonton: University of Alberta Press, 81-91.
- Adapte la méthode de l'auteur, pour enquêtes avec multiplicité pour des événements rares, au problème du système dual.
- SMITH, P.J. (1988). Bayesian methods for capture-recapture surveys. *Biometrics*, 44, 1177-1189.
- Utilise l'approximation de Poisson et la distribution gamma à priori pour une méthode bayésienne appliquée à l'estimation quand il y a indépendance dans un modèle à resaisies multiples.
- SMITH, P.J. (1991). Bayesian analyses for multiple capture-recapture model. *Biometrika*, 78, 399-407.
- Élabore une distribution bayésienne *a posteriori* exacte pour un recensement à resaisies multiples quand les resaisies sont indépendantes.
- SRINIVASAN, S.K., et MUTHIAH, S.A. (1968). Problems of matching births identified from two independent sources. *Journal of Family Welfare*, 14, 13-22.
- TRACY, W.R. (1941). Fécondité de la femme canadienne. Réimprimé de *Septième recensement du Canada*, 1931, (Vol. 2), Monographies du recensement n° 3. Ottawa: Cloutier.
- Une des premières applications de la méthode de système dual aux données du recensement.
- WINKLER, W.E. (1989). Méthodes permettant de tenir compte de l'absence d'indépendance dans une application du modèle d'appariement des enregistrements de Fellegi-Sunter. *Techniques d'enquête*, 15, 105-122.
- Examine des méthodes pour ajuster les règles d'appariement pour des enregistrements de système dual apparés quand on ne peut supposer qu'il y a indépendance.

- RAJ, D. (1977). On estimating the number of vital events in demographic surveys. *Journal of the American Statistical Association*, 72, 377-381.
- Elabore une formule pour calculer le biais dans l'estimateur de système dual pour un modèle général des erreurs de réponse et explore l'utilisation de l'échantillonnage double pour corriger le biais.
- ROSSMO, D.K., et ROUTLEDGE, R. (1990). Estimating the size of criminal populations. *Journal of Quantitative Criminology*, 6, 293-314.
- RUBIN, D.B., SCHAFER, J.L., et SCHENKER, N. (1988). Méthodes d'imputation de valeurs manquantes dans des enquêtes postcensitaires. *Techniques d'enquête*, 14, 223-236.
- Présente une méthode d'appariement pour l'estimation du sous-dénombrement qui utilise une technique d'imputation tirant son origine dans les modèles log-linéaires quand des données manquent.
- SANATHANAN, L.P. (1972a). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- Démonstre l'équivalence asymptotique d'estimateurs conditionnels et inconditionnels de la taille d'une population.
- SANATHANAN, L.P. (1972b). Models and estimation methods in visual scanning experiments. *Technometrics*, 14, 813-829.
- Elabore un modèle latent afin d'estimer le nombre de particules dont on a besoin pour balayer des enregistrements, modèle qui tient compte de la détectabilité différentielle et qui engendre des dépendances parmi les détecteurs.
- SANATHANAN, L.P. (1973). A comparison of some models in visual scanning experiments. *Technometrics*, 15, 67-78.
- Applique le modèle traditionnel de saisie-resaisie et des modèles latents à des données provenant d'expériences réelles de balayage dans le domaine visuel.
- SANDLAND, R.L., et CORMACK, R.M. (1984). Statistical inference for Poisson and multinomial models for capture recapture experiments. *Biometrika*, 71, 27-33.
- Montre le rapport entre les variances asymptotiques de la taille de la population dans un modèle général de saisie-resaisie pour les deux plans d'échantillonnage qui peuvent être utilisés.
- SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14, 93-104.
- Examine l'effet de données manquantes sur l'estimation de système dual appliquée à des données de recensement d'essai.
- SCHIRM, A.L., et PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (avec discussion). *Journal of the American Statistical Association*, 82, 965-990.
- Applique des méthodes d'estimation synthétique aux données du recensement de 1980 pour évaluer l'incidence de l'estimation du sous-dénombrement.
- SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.
- Etend la méthode de base de saisie-resaisie à des resaisies multiples, avec, lors de chaque resaisie, des renseignements qui disent si les individus avaient été saisis auparavant.
- SCOTT, C. (1974). The dual record (PGE) system for vital rate measurement, some suggestions for further development. Dans *Congrès international de la population, Liège, 1973*, (Volume 2). Liège: Union internationale pour l'étude scientifique de la population.

- LEWIS, C.E., et HASSANEIN, K.M. (1969). The relative effectiveness of different approaches to the surveillance of infection among hospitalized patients. *Medical Care*, 8, 379-384.
- Applique l'estimation de système dual à la surveillance des maladies infectieuses.
- LINCOLN, F.C. (1930). Calculating waterfowl abundance on the basis of banding returns. *Circular of the U.S. Department of Agriculture*, 118, 1-4.
- Applique la méthode de saisie-resaisie à l'estimation de la taille des populations d'oiseaux aquatiques.
- MANTEL, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics*, 7, 240-246.
- Montre comment l'hétérogénéité entraîne un biais de corrélation (corrélation des événements) dans l'estimation de la prévalence des maladies.
- MARKS, E.S., SELTZER, W., et KRÖTKI, K.J. (1974). *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: Population Council.
- Examen complet des hypothèses, des données de base, de la conception et des problèmes relatifs à l'estimation de système dual. Renferme une affirmation selon laquelle la méthode de base est utilisée depuis plus de trois siècles pour estimer la taille de populations animales.
- MAXIM, L.D., HARRINGTON, L., et KENNEDY, M. (1981). A capture-recapture approach for estimation of detection probabilities in aerial surveys. *Photogrammetric Engineering and Remote Sensing*, 47, 779-788.
- MULRY, M.H., et SPENCER, B.D. (1988). L'erreur totale dans l'estimateur de système dual: Recensement du Central Los Angeles County de 1986. *Techniques d'enquête*, 14, 257-280.
- Elabore un modèle de l'erreur totale pour la méthode de système dual appliquée au test des opérations de redressement de Los Angeles.
- MULRY, M.H., et SPENCER, B.D. (1991). Total error in PES estimates of population (avec discussion). *Journal of the American Statistical Association*, 86, 839-863.
- Étend le développement antérieur de Mulry-Spencer du modèle de l'erreur totale pour la méthode de système dual et l'applique à la répétition générale du recensement qui a eu lieu en 1988 à Saint Louis et dans le centre est du Missouri.
- NICHOLS, J.D., et POLLOCK, K.H. (1983). Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture-recapture models. *Paleobiology*, 9, 150-163.
- OTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monograph*, 62, Washington, DC: Wildlife Society.
- Passe en revue la méthode de saisie-resaisie et les méthodes connexes pour des populations fauniques.
- PERKINS, W.M., et JONES, C.D. (1965). Matching for census coverage checks. Communication présentée lors de la réunion de l'American Statistical Association, Philadelphie.
- PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station to the Ministry of Fisheries*, 6, 1-48.
- Elabore la méthode de saisie-resaisie et de son application à l'estimation de la taille de populations de poissons.
- POLLACK, E.S. (1965). Use of census matching for study of psychiatric admission rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-115.

- ISAKI, C.T., et SCHULTZ, L.K. (1987). The effect of correlation and matching error in dual system estimation. *Communications in Statistics, Theory and Methods*, 16, 2405-2427.
- Elabore un modèle d'erreur d'appariement simple en présence d'un biais de corrélation pour compenser trois estimateurs de système dual.
- ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., et HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.
- Elabore des populations de simulation basées sur les données du recensement de 1980 et sur les résultats d'une évaluation de la couverture, évalue des méthodes synthétiques d'estimation du sous-dénombrement basées sur la régression et montre la supériorité des méthodes synthétiques par rapport aux chiffres bruts du recensement.
- JABINE, T.B., et BERSHAD, M.A. (1968). Some comments on the Chandrasekar and Deming technique for the measurement of population change. Communication présentée lors du CENITO Symposium on Demographic Statistics, Karachi, Pakistan.
- Montre qu'un biais de corrélation positif produit un biais par défaut dans l'estimation de la taille de la population totale.
- JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Test Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- Décrit une méthodologie du recensement pour appairer des enregistrements du recensement avec ceux de l'enquête post-censitaire, présente aussi les résultats tirés de l'application de cette méthodologie au recensement d'essai de 1985.
- JELINSKI, Z., et MORANDA, P.B. (1972). Software reliability research. Dans *Statistical Computer Performance Evaluation*, (éd. W. Freiberger). New York: Academic Press, 465-484.
- Propose un modèle avec incidents distribués exponentiellement pour estimer le nombre total d'anomalies dans un programme en se basant sur le moment où les incidents se produisent au cours d'une période de durée fixe.
- JEWELL, W.S. (1985). Bayesian estimation of undetected errors. Dans *Bayesian Statistics 2*, (éds. J.M. Bernardo, et coll.). New York: Elsevier. 663-671.
- JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration - stochastic models. *Biometrika*, 52, 225-247.
- Estimation à partir de données sur des resaisies multiples pour des populations ouvertes.
- KADANE, J.B., MEYER, M.M., et TUKEY, J.W. (1992). Correlation bias in the presence of stratum heterogeneity. Soumis pour publication.
- Démontre l'incidence du biais de corrélation découlant d'un regroupement parmi des strates hétérogènes avec des probabilités différentes du potentiel de capture dans chaque strate, quand il y a une contrainte monotone.
- KRÖTKI, K.J. (éd.) (1978). *Developments in Dual System Estimation of Population Size and Growth*. Edmonton: University of Alberta Press.
- Examine l'utilisation de l'estimation de système dual pour les actes de l'état civil dans divers pays. Comprend des détails techniques sur l'utilisation d'échantillons complexes et des élaborations sur les techniques de base.
- LASKA, E.M., MEISNER, M., et SIEGEL, C. (1988). Estimating the size of a population from a single sample. *Biometrics*, 44, 461-472. Correction, (1989), 45, 1347.
- Estime la taille de la population à partir de la dernière de k listes.

- GREENFIELD, C.C. (1975). On the estimation of a missing cell in a 2×2 contingency table. *Journal of the Royal Statistical Society, Série A*, 138, 51-61.
- Introduit une valeur non nulle pour la corrélation de réponses en prenant le point milieu de la gamme de valeurs de corrélation permises et obtient, par conséquent, une valeur pour une case manquante. Applique la méthode à des données du recensement du Malawi.
- GREENFIELD, C.C. (1976). A revised procedure for dual record systems in estimating vital events. *Journal of the Royal Statistical Society, Série A*, 139, 389-401.
- Applique des limites à la corrélation dans un tableau 2×2 à l'estimation de système dual en présence de corrélations entre des événements engendrés par l'hétérogénéité.
- GREENFIELD, C.C., et TAM, S.M. (1976). A simple approximation for the upper limit to the value of a missing cell in a 2×2 contingency table. *Journal of the Royal Statistical Society, Série A*, 139, 96-103.
- Utilise une approximation de la limite supérieure de la corrélation de réponse afin d'obtenir une limite supérieure pour une case manquante.
- HOGAN, H., et WOLTER, K.M. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.
- Rend compte du test des opérations de redressement mené à Los Angeles et estime les sources de biais dans une enquête post-censitaire et dans les estimations de systèmes dual basées sur les données d'un recensement.
- HOLST, L. (1973). Some limit theorems with applications in sampling theory. *Annals of Statistics*, 1, 644-658.
- Applique les résultats d'un échantillonnage successif pour obtenir une distribution asymptotique de l'estimateur habituel de Peterson quand il y a des probabilités de saisie hétérogènes ou des effets d'appariement.
- HOOK, E., et REGAL, R. (1982). Validity of Bernoulli census, log-linear, and truncated binomial models for correcting underestimates in prevalence studies. *American Journal of Epidemiology*, 116, 168-176.
- Applique des méthodes différentes liées aux logarithmes linéaires utilisés pour étudier le nombre d'enfants nés avec le syndrome de Down.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- Utilise des modèles logistiques linéaires pour des probabilités de saisie pour des individus et des occasions de saisie.
- HUGGINS, R.M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47, 725-732.
- Utilise des modèles logistiques linéaires pour des probabilités de saisie et exploite l'ordonnement temporel des saisies pour introduire une dépendance parmi les saisies et parmi les covariances mesurables pour les individus qui ont été saisis au moins une fois.
- ISAKI, C.T. (1986). Bias of the dual system estimator and some alternatives. *Communications in Statistics, Theory and Methods*, 15, 1435-1450.
- Exploite une limite supérieure du biais de corrélation afin de réduire le biais de l'estimateur de système dual.
- ISAKI, C.T., et SCHULTZ, L.K. (1986). Dual system estimation using demographic analysis data. *Journal of Official Statistics*, 2, 169-179.
- Utilise des données produites par analyse démographique pour obtenir des estimations de système dual révisées pour le recensement de 1980 à l'aide de modèles différents du biais de corrélation.

FAY, R.E., PASSSEL, J.S., ROBINSON, J.G., et COWAN, C.D. (1988). *The Coverage of Population in the 1980 Census*. Bureau of the Census. Washington D.C.: U.S. Department of Commerce.

• Rapport officiel du Bureau of the Census sur les tentatives visant à mesurer le sous-dénombrement du recensement décennal des E.-U. de 1980.

FEIN, D.J., et WEST, K.K. (1988). Sources du sous-dénombrement lors du recensement: Résultats du recensement d'essai de 1986 à Los Angeles. *Techniques d'enquête*, 14, 237-256.

• Tente de faire des tests d'hypothèses en rapport avec les causes du sous-dénombrement du recensement pour une population urbaine composée de personnes d'origine hispanique difficiles à dénombrer.

FIENBERG, S.E. (1972). The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika*, 59, 591-603.

• Présente une méthode pour estimer les dépendances parmi des listes multiples à l'aide de modèles log-linéaires et élabore une méthode générale d'estimation, applique les résultats de l'élaboration de cette méthode à des tableaux de contingence incomplets 2^k et à une estimation conditionnelle.

FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. Dans *Encyclopedia of Statistical Sciences*, (Supplément), (éds. S. Kotz et N.L. Johnson), New York: Wiley, 181-185.

• Présente un aperçu historique de la différence du taux de sous-dénombrement de la population des E.-U. ainsi que de brèves descriptions des méthodes d'estimation de système dual et à l'aide de l'analyse démographique.

FREEEDMAN, D.A. (1991). Policy forum: Adjusting the 1990 census. *Science*, 252, 1233-1236.

• Critique de la méthode de système dual pour redresser les chiffres du recensement de 1990.

FREEEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models and adjusting the 1980 census (avec discussion). *Statistical Science*, 1, 3-39.

• Critique de la méthodologie de système dual d'Ericsson et Kadane telle qu'appliquée aux données du recensement de 1980.

FREEEDMAN, D.A., et NAVIDI, W.C. (1992). Aurions-nous dû redresser les chiffres du recensement de 1980? (avec discussion). *Techniques d'enquête*, dans le présent numéro.

• Poursuit la critique de l'utilisation de l'estimation de système dual et du redressement synthétique tels qu'appliqués aux chiffres du recensement de 1980.

GARTHWAITE, P.H., et BUCKLAND, S.T. (1990). Analysis of multiple-recapture census by computing conditional probabilities. *Biometrics*, 46, 231-238.

• Utilise un rapport récursif pour produire des estimations ponctuelles et par intervalle pour un recensement par resaisies multiples quand il y a indépendance.

GEIGER, H., et WERNER, A. (1924). Die Zahl der ion radium ausgesandten a-Teilchen. *Zeitschrift für Physik*, 21, 187-203.

• Applique la méthode de saisie-resaisie à une estimation du nombre d'ions radium détectés.

GOLDBERG, J.D., et WITTES, J.T. (1978). The estimation of false negatives in medical screening. *Biometrics*, 34, 77-86.

• Applique des modèles de saisie-resaisie à des problèmes de dépistage.

GOUDIE, I.B.J. (1990). A likelihood-based stopping rule for recapture debugging software reliability. *Biometrika*, 77, 203-206.

GREEN, M.A., et STOLLMACK, S. (1981). Estimating the number of criminals. Dans *Models in Quantitative Criminology*, (éd. J.A. Fox), New York: Academic Press, 1-24.

- DAVIDSON, L. (1962). Retrieval of misspelled names in an airline passenger record system. *Communications of the Association of Computer Machinery*, 5, 169-171.
- DEMING, W.E., et KEYFITZ, N. (1967). Theory of surveys to estimate total populations. Dans *Compte rendu de la Conférence mondiale sur la population*, Belgrade, 1965, (Vol. 3). New York: Nations Unies, 141-144.
- Étend la méthode de Chandrasekar-Deming à trois sources.
- DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14, 75-92.
- Décrit la mise en application de la méthode d'enquête postcensitaire à l'estimation de système dual dans un recensement d'essai.
- DING, Y. (1990). Capture-Recapture Census with Uncertain Matching. Thèse de doctorat, Département de statistique, Carnegie Mellon University.
- Élabore un modèle d'appariement probabiliste pour utilisation avec une estimation de système dual et de système multiple et considère une méthode bayésienne pour estimer la taille de la population. Illustre les techniques à l'aide de données provenant des résultats d'un recensement d'essai effectué à Los Angeles.
- DING, Y., et FIENBERG, S.E. (1992). Estimating population and census undercount in the presence of matching error. Soumis pour publication.
- Élabore un modèle d'appariement probabiliste pour utilisation avec une estimation de système dual et en illustre l'application à des données provenant des résultats d'un recensement d'essai effectué à Los Angeles.
- DURAN, J.W., et WIOORKOWSKI, J.J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Systems Engineering*, 7, 147-148.
- EFRON, B., et THISTED, R.A. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63, 435-467.
- Adapte un modèle paramétrique attribué à Fisher et un modèle non paramétrique au problème classique des espèces à l'aide de méthodes empiriques de Bayes. Applique la méthode au vocabulaire de Shakespeare.
- EL-KHORAZATY, M.N., IMREY, P.B., KOCH, G.G., et WELLS, H.B. (1977). Estimating the total number of events with data from multiple record systems: a review of methodological strategies. *International Statistical Review*, 45, 129-157.
- Passe en revue les ouvrages et les méthodes relatives à l'estimation de système dual et de système multiple. Comprend des sections où l'on compare l'utilisation de techniques et des situations où l'on s'éloigne des hypothèses, dans les populations fauniques et humaines.
- ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.
- Applique la méthode de système dual aux données du recensement de 1980, y compris le lissage, par la méthode de régression, des estimations du sous-dénombrement ainsi que l'estimation redressée, à l'aide d'estimations démographiques, du risque relatif.
- ERICKSEN, E.P., KADANE, J.B., et TURKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.
- Présente des révisions et des prolongements de la méthodologie d'Ericksen et Kadane et une critique de Freedman et Navidi.

- CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- Elabore et applique des méthodes empiriques de lissage de Bayes pour des facteurs de redressement des chiffres du recensement produits à l'aide de la méthode de système dual pour stratification en fonction de la géographie et de la démographie. Applique la méthode aux données sur les Etats tirées du recensement des E.-U. de 1980.
- CRESSIE, N., et DAJANI, A. (1991). Empirical Bayes estimation of U.S. undercount based on artificial populations. *Journal of Official Statistics*, 7, 57-67.
- Montre que la méthode d'estimation synthétique utilisée par Isaki et coll. est un cas spécial de la méthode empirique de Bayes.
- CROXFORD, A.A. (1968). Record linkage in education. Dans *Record Linkage in Medicine* (éd. E.D. Acheson). Londres: E. & S. Livingstone, 351-356.
- DAHIVA, R.C., et BLUMENTHAL, S. (1986). Population or sample size estimation. Dans *Encyclopedia of Statistical Sciences*, (Volume 7), (éds. S. Kotz et N.L. Johnson). New York: Wiley, 100-110.
- Passe en revue la théorie qui est à la base de l'estimation de la taille d'une population à partir d'un échantillonnage tronqué pour une distribution discontinue et fournit des références pour des domaines d'application.
- DARROCH, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.
- Décrit la méthode du maximum de vraisemblance appliquée au problème de la resaisie multiple quand il y a indépendance complète.
- DARROCH, J.N. (1959). The multiple-recapture census II: Estimation when there is immigration or death. *Biometrika*, 46, 336-351.
- Étend la méthode du maximum de vraisemblance quand il y a indépendance à des populations ouvertes avec immigration ou décès.
- DARROCH, J.N. (1961). The two sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, 45, 343-359.
- Étend la méthode du maximum de vraisemblance avec indépendance à la situation où les individus saisis à l'origine sont stratifiés en s groupes et où les individus dans l'échantillon de resaisie sont stratifiés, mais selon t strates (possiblement différentes).
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., et JUNKER, B.W. (1992). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. Soumis pour publication.
- Étend l'estimation de système triple afin de tenir compte de l'hétérogénéité des individus et de formes choisies de dépendance. Applique les estimateurs aux données de système triple provenant de la répétition générale du recensement qui s'est tenue à Saint Louis.
- DARROCH, J.N., et RATCLIFF, D. (1980). A note on capture-recapture estimation. *Biometrics*, 36, 149-153.
- Présente un autre estimateur pour les problèmes de saisie-resaisie avec des propriétés asymptotiques intéressantes.
- DASGUPTA, P. (1964). On the estimation of the total number of events and of the probabilities of detecting an event from information supplied by several agencies. *Calcutta Statistical Association Bulletin*, 13, 89-100.
- Étend la méthode de Chandrasekar-Deming à trois sources ou plus.

- Elabore le modèle d'échantillonnage hypergéométrique pour estimer la taille de la population dans des études de saisie-resaisie.
- CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.
- Décrivit l'utilisation de l'estimation de système dual et d'une enquête postcensitaire pour redresser les chiffres du recensement de l'Australie. Appliqua aussi la technique du rapport de masculinité de Wolter pour vérifier la sensibilité de l'estimateur de système dual.
- CHRISTENSEN, H.T. (1958). The method of record linkage applied to family data. *Marriage and Family Living*, 20, 38-43.
- CITRO, C.F., et COHEN, M.L., (éds.) (1985). *The Bicentennial Census. New Directions for Methodology in 1990*. Washington D.C.: National Academy Press.
- Rapport d'un groupe d'experts du Committee on National Statistics sur la méthodologie du recensement, le rapport comprend un examen de la méthode de système dual pour corriger le sous-dénombrement.
- COALE, A.J. (1961). The design of an experimental procedure for obtaining accurate vital statistics. *International Population Conference*, New York, 372-375.
- Propose d'utiliser deux listes qui se rapportent au même échantillon d'une population.
- COHEN, M.L. (1990). Adjustment and reapportionment - analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.
- Examine l'effet du biais et de la variabilité sur la précision des chiffres redressés et non redressés du recensement ainsi que l'incidence de la nouvelle répartition des sièges à la Chambre des représentants des E.-U.
- CORMACK, R.M. (1981). Loglinear models for capture-recapture experiments on open populations. Dans *The Mathematical Theory of the Dynamics of Biological Populations*, II (éds. R.W. Hiorns et D. Cooke). Londres: Academic Press, 217-235.
- Présente le modèle de Poisson pour la saisie-resaisie et utilise ce dernier avec des modèles log-linéaires pour étendre la méthode standard afin de tenir compte de la naissance, du décès et de la dépendance par rapport aux pièges.
- CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.
- Utilise le modèle de Poisson et une représentation log-linéaire pour inclure la naissance, le décès et la dépendance par rapport aux pièges dans la méthode standard de saisie-resaisie.
- CORMACK, R.M., et JUPP, P.E. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, 78, 911-916.
- Compare les estimateurs du maximum de vraisemblance de paramètres selon deux modèles et présente le rapport entre les variances et covariances asymptotiques correspondantes.
- COWAN, C.D., et MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.
- Étend la méthode de système dual à une situation comportant des observations groupées comme dans le programme d'amélioration de la couverture du recensement des E.-U.
- CRESSIE, N. (1988). Dans quelles circonstances les opérations de redressement améliorent-elles les chiffres du recensement? *Techniques d'enquête*, 14, 205-222.
- Propose un modèle basé sur l'enquête postcensitaire pour corriger le sous-dénombrement à l'aide d'un plan d'estimation empirique de Bayes et d'une famille de fonctions de perte.

- BOSWELL, M. T., BURNHAM, K. P., et PATIL, G. P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies. Dans *Handbook of Statistics 6: Sampling* (éd. P. R. Krishnamiah et C. R. Rao). Amsterdam: North Holland, 469-488.
- Présente un résumé succinct de plusieurs variantes de base des modèles de saisie-resaisie ainsi que de leur estimation.
- BROWNIE, C., ANDERSON, D. R., BURNHAM, K. P., et ROBSON, D. S. (1977). Statistical inference from band recovery data: a handbook. *U.S. Fisheries and Wildlife Service Resource Publication No. 131*.
- Décrit une gamme étendue de modèles de saisie-resaisie et les tests appropriés de validité de l'ajustement, avec l'accent mis sur des expériences de marquage et de baguage.
- BURGESS, R. D. (1988). Evaluation des estimations du sous-dénombrement obtenues par la contre-vérification des dossiers du recensement du Canada, *Techniques d'enquête*, 14, 147-167.
- Décrit la méthode "comptable" basée sur une enquête employée pour effectuer la contre-vérification des dossiers afin d'estimer le sous-dénombrement. Ne traite pas de la question de l'exclusion de personnes du recensement et d'autres listes.
- BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., et POLLOCK, K. H. (1987). *Design and Analysis Methods for Fish Survival Experiments Based on Release-Recapture*. Bethesda, MD: American Fisheries Society.
- Combine la méthodologie de Brownie et coll. pour le recouvrement de marques avec estimation de la survie selon les modèles de marquage et recapture de Jolly-Seber.
- BURNHAM, K. P., et OVERTON, W. S. (1978). Estimation of the size of a closed population when the capture probabilities vary among animals. *Biometrika*, 65, 625-633. Correction (1981) 68, 345.
- Élabore un modèle de saisie-resaisie avec hétérogénéité pour les animaux, mais probabilités constantes de saisie dans tous les échantillons. Le modèle entraîne des dépendances parmi les captures.
- CASTELDINE, B. J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.
- Élabore une méthode bayésienne à l'aide de la distribution bêta à priori pour un modèle de recensement traditionnel de Schnabel basé sur l'indépendance pour des données sur des saisies multiples.
- CHAKRABORTY, P. N. (1963). On a method of estimating birth and death rates from several agencies. *Calcutta Statistical Association, Bulletin*, 12, 106-112.
- Étend la méthode de Chandrasekar-Deming à trois sources ou plus.
- CHANDRASEKAR, C., et DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- Élabore une technique de système dual et propose d'utiliser la stratification pour éliminer l'hétérogénéité. Applique la méthode pour estimer le nombre de naissances et de décès dans plusieurs villages de l'Inde.
- CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43, 783-791.
- Étudie le modèle hétérogène du potentiel de capture de Burnham et Overton à l'aide d'une inégalité de moments pour obtenir une limite inférieure de la taille de population.
- CHAO, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45, 427-438.
- Étudie la suffisance de l'estimateur obtenu à partir d'une inégalité des moments pour un modèle hétérogène du potentiel de capture dans des situations qui comprennent l'utilisation de données éparpillées.

recensement pour évaluer la couverture. Cette méthode a évolué pour devenir ce que nous connaissons maintenant comme la méthode de l'enquête post-censitaire pour estimer le sous-dénombrement et le surdénombrement et elle a été le point central de la controverse récente et qui se poursuit sur le redressement possible des chiffres des recensements de 1980 et de 1990 (p. ex., voir Erikssen et Kadane 1985; Freedman et Navidi 1986, 1992; Freedman 1991; Wolter 1991).

On trouve dans la présente bibliographie commentée sommaire un aperçu des ouvrages publiés sur l'estimation des totaux de population à l'aide de la méthode de saisie-resaisie. La bibliographie comprend des références historiques, des articles qui explorent des situations où l'on s'écarte des hypothèses et des prolongements de la méthodologie de base et c'est pour les articles qui décrivent les méthodes de système dual et de système multiple dans le contexte de l'estimation du sous-dénombrement du recensement qu'elle est la plus complète. À ce sujet, toutefois, nous n'avons pas inclus de référence à l'un quelconque des mémoires et documents non publiés du U.S. Bureau of Census (surtout parce que la majorité de ces documents ont été repris sous une certaine forme dans les ouvrages publiés). Nous avons eu tendance à exclure les articles qui ont paru dans des publications sans comité de lecture pour des raisons connexes. Comme les ouvrages sur les applications spécialisées des techniques de saisie-resaisie aux populations fauniques sont très nombreux et que seulement certains d'entre eux s'appliquent aux populations humaines, nous avons surtout fourni des références à des comptes rendus de ces ouvrages, p. ex., voir Brownie et coll. 1977; Otis et coll. 1978; Seber 1973, 1982. De même, nous n'avons inclus qu'un petit nombre de références aux méthodes plus spécialisées utilisées pour mesurer la durée de vie, p. ex., voir Dahya et Blumenthal 1986, ainsi que celles qui sont employées dans des applications relatives à la fiabilité du logiciel, p. ex., Jelinski et Moranda 1972, et Duran et Wiorowski 1981. Les méthodes utilisées dans ces derniers ouvrages divergent considérablement de celles employées dans les méthodes de base de saisie-resaisie et de système dual.

2. BIBLIOGRAPHIE SOMMAIRE

- ALHO, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- Prolonge la méthode habituelle de système dual afin de pouvoir tenir compte des effets multiplicatifs de la stratification.
- BAKER, S. G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates (avec discussion). *Biometrics*, 46, 1193-1200.
- Lie le classement recoupé de covariables à la saisie et à la resaisie par l'intermédiaire de modèles log-linéaires puis utilise l'algorithme EM pour estimer la taille de la population.
- BIEMER, P. P. (1988). Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement. *Techniques d'enquête*, 14, 125-143.
- Elabore des modèles pour évaluer l'incidence de l'erreur d'appariement sur la couverture du recensement.
- BISHOP, Y. M. M., FIENBERG, S. E., et HOLLAND, P. H. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Chapitre 6. Cambridge, MA: MIT Press.
- Monographie sur les modèles log-linéaires qui comprend un chapitre sur le rapport avec les modèles de saisie-resaisie.
- BLUMENTHAL, S., et MARCUS, R. (1975). Estimating population size with exponential failure. *Journal of the American Statistical Association*, 70, 913-922.
- Utilise une distribution exponentielle pour estimer la taille de la population à partir d'un sous-ensemble d'observations obtenues par échantillonnage tronqué.

Bibliographie sur la modélisation à l'aide de la saisie-resaisie avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement

STEPHEN E. FIENBERG¹

RÉSUMÉ

Dans cet article on présente un choix bibliographique commenté d'ouvrages sur l'estimation de la taille de la population par saisie-resaisie (système dual), sur des prolongements de la méthodologie de base et sur l'application de ces techniques dans le contexte de l'estimation du sous-dénombrement du recensement.

MOTS CLÉS: Saisie-resaisie; sous-dénombrement du recensement; estimation de système dual; modèles log-linéaires.

1. INTRODUCTION

La méthode de la saisie-resaisie pour estimer la taille d'une population fermée est utilisée depuis au moins le dix-neuvième siècle, quand Peterson (1896) a élaboré l'estimateur standard qui porte son nom pour utilisation avec des populations de poissons. Parmi les applications faites par la suite à d'autres genres de populations notons: Geiger et Werner (1924) – pour des études en physique; Lincoln (1930) – pour étudier la faune; Chandrasekar et Deming (1948) – pour étudier les statistiques de l'état civil pour des populations humaines; Wittes et Sidel (1968), Wittes, Colton et Sidel (1974) – pour des études en épidémiologie; Sanathanan (1972b) – pour étudier le balayage par faisceaux de particules en physique; Blumenthal et Marcus (1975) – pour mesurer la durée de vie; Green et Stollmack (1981), Rossmo et Rouledge (1990) – pour étudier les crimes et les criminels. Dans le contexte de l'étude des populations humaines et de la démographie, on désigne souvent cette méthode par l'expression "estimation de système dual". Nous n'avons inclus presque aucune référence au problème connexe qui consiste à dénombrer le nombre d'espèces, problème qui remonte au travail de R. A. Fisher au cours des années 40 et qui est formulé de façon élégante dans l'article de Efron et Thisted (1976) paru dans *Biometrika* et intitulé "How many words did Shakespeare know?".

La méthode de saisie-resaisie de base est fondée sur un certain nombre d'hypothèses, p. ex.: (1) la population étudiée est fermée; (2) on peut apparter parfaitement les individus (unités) de la saisie à la resaisie; (3) les probabilités de saisie sont constantes pour tous les individus (unités) dans la population; (4) la probabilité qu'un individu (unité) soit inclus dans l'échantillon utilisé pour la resaisie est indépendante de l'inclusion de l'individu (de l'unité) dans le recensement ou l'échantillon original. À compter de la fin des années 30, divers chercheurs ont commencé à étudier des prolongements de la méthode qui permettaient de s'écarter des hypothèses. Ces techniques requièrent généralement des données additionnelles comme une deuxième resaisie (ou même une troisième) et les données complètes sur les saisies-resaisies pour chaque individu.

Pour les populations humaines et l'étude des statistiques de l'état civil, la méthodologie a, depuis longtemps, été liée aux données du recensement, p. ex., voir Tracy (1941) et Shapiro (1949, 1954). À l'occasion du recensement décennal de la population de 1950, le U.S. Bureau of the Census a introduit l'utilisation d'un échantillon apparié aux enregistrements du

¹ Stephen E. Fienberg, Office of Vice President (Academic Affairs), Université York, North York (Ontario) M3J 1P3, Canada.

STATISTIQUE CANADA (1988). Area Master File (AMF), User guide. Rapport interne, Statistique Canada.

STATISTIQUE CANADA (1989a). Automated Postal Coding System (PCODE), User and retrieval guide. Rapport interne, Statistique Canada.

STATISTIQUE CANADA (1989b). Generalized Iterative Record Linkage System, Concepts guide. Rapport interne, Statistique Canada.

STATISTIQUE CANADA (1989c). Postal Address Analysis System (PAAS), User guide. Rapport interne, Statistique Canada.

STATISTIQUE CANADA (1989d). Record linkage software, Reference guide. Rapport interne, Statistique Canada.

STATISTIQUE CANADA (1990). *Guide à l'intention des utilisateurs sur la qualité des données du recensement de 1986: Couverture*. 99-135F, Statistique Canada.

STATISTIQUE CANADA (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide. Rapport interne, Statistique Canada.

SWAIN, L., DREW, J.D., LAFRANCE, B., et LANCE, K. (1992). La création d'un registre d'adresses résidentielles à Statistique Canada. *Recueil du Symposium sur les questions spatiales liées aux statistiques*. Statistique Canada.

VAN BAAREN, A. (1988). Report on the November 1987 address register test. Rapport interne, Statistique Canada.

sources ont été fournis à Statistique Canada à titre confidentiel, soit dans le cadre d'un contrat (p. ex., certains fichiers de l'Alberta), soit en vertu de la loi (le fichier des déclarations de revenus des particuliers (T1) de Revenu Canada).

6.7 Conclusion

L'ampleur et la diversité des idées mentionnées plus haut dans les orientations futures démontrent les possibilités qu'offre le registre des adresses comme produit géographique avec des applications dans de nombreux domaines qui intéressent Statistique Canada et des organismes de l'extérieur.

REMERCIEMENTS

Les auteurs désirent remercier les nombreuses personnes qui font partie des services mentionnés ci-après pour leur dévouement et leur persévérance lors de la création du registre des adresses: Phillip Reed ainsi que la sous-section de la production du RA, Division de la géographie; l'unité du contrôle de l'échantillon de l'enquête sur la population active, Méthodes de recensement, Division des opérations des enquêtes; le Centre principal des ordinateurs et la Division des enquêtes des ménages. Les auteurs désirent aussi remercier l'arbitre, Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed et Carol Sol pour leurs suggestions utiles lors de la réalisation du présent article.

BIBLIOGRAPHIE

- BOOTH, J.K. (1976). A summary report of all address register studies completed to date. Rapport interne, Statistique Canada.
- DICK, P. (1990). Address register – September 1989 test. Ébauche interne, Statistique Canada.
- DREW, J.D., ARMSTRONG, J.B., et DIBBS, R. (1987). Research into a register of residential addresses for urban areas of Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 300-305.
- DREW, J.D., ARMSTRONG, J., VAN BAREN, A., et DEGUIRE, Y. (1988). Méthodologie de la constitution d'un registre d'adresses à partir de plusieurs sources administratives. *Recueil du Symposium sur les utilisations statistiques des données administratives*, Statistique Canada, 209-219.
- FELLEGI, I.P., et KRÖTKI, K.J. (1967). The testing program for the 1971 Census in Canada. *Proceedings of the Social Statistics Section, American Statistical Association*, 29-38.
- FELLEGI, I.P., et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- GAMACHE-O'LEARY, V., NIEMAN, L., et DIBBS, R. (1987). Cost implications of mail-out of Census questionnaires using an address register. Rapport interne, Statistique Canada.
- HILL, T., et PRING-MILL, F. (1985). Generalized iterative record linkage system. *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.
- ROYCE, D. (1986). Address register research for the 1991 Census of Canada. *Journal of Official Statistics*, 2, 4, 447-455.
- ROYCE, D. (1987). Applications du registre des adresses au recensement du Canada. *Actes de la conférence internationale sur la planification du recensement de 1991*, Statistique Canada, 227-237.
- ROYCE, D., et DREW, J.D. (1988). Address register research: Current status and future plans. Rapport interne, Statistique Canada.

6.4 Méthodologie de mise à jour

Dans une large mesure, on a déjà épuré les systèmes informatiques élaborés pour la production initiale du RA afin d'augmenter l'efficacité de l'utilisation du gros ordinateur, des programmes, du stockage sur disques et sur bandes, de la manipulation des fichiers, des sorties, de l'accès aux bibliothèques ainsi qu'aux fichiers. De meilleurs contrôles des systèmes seront préparés. Ce RA n'a été produit que pour les régions urbaines. Au cours d'un développement méthodologique ultérieur, on examinera la possibilité de l'étendre aux régions rurales.

Le RA a été créé à partir de quatre ensembles de fichiers administratifs: les fichiers des sociétés de téléphone, les fichiers des rôles d'évaluation des municipalités, les fichiers des sociétés d'électricité ainsi que le fichier de déclaration de revenu des particuliers (T1) de Revenu Canada. De plus, la mise à jour du RA est actuellement en cours afin que ce dernier soit compatible avec le recensement de 1991, de sorte que le recensement est aussi une source de données. Les contributions relatives de ces fichiers sources, tant en volume qu'en qualité, seront étudiées afin qu'on puisse prendre une décision pour ce qui est de l'acquisition de fichiers en vue d'effectuer la mise à jour.

L'élaboration d'une méthodologie pour effectuer la mise à jour constitue une partie intégrante de la stratégie de mise à jour. On aura besoin de la définition d'une mise à jour ainsi que d'un système de mise à jour. On considérera aussi le rapport coût-efficacité d'une mise à jour continue, selon les divers besoins qui découlent des projets définis dans toutes ces orientations futures. La mise à jour continue est-elle rentable quand on la compare à la mise à jour effectuée seulement à temps pour le recensement? Quelles exigences découleront d'autres utilisations possibles? Les réponses à ces questions nous amèneront à élaborer une stratégie de mise à jour.

6.5 Autres utilisations du registre des adresses à Statistique Canada

En plus du recensement et des rapports géographiques présentés plus haut, un certain nombre d'autres utilisations sont proposées à l'intérieur de Statistique Canada. L'utilisation possible du RA pour l'enquête sur la population active (EPA) sera étudiée dans le cadre du projet de remaniement de l'EPA. La possibilité d'utiliser le RA dans les régions urbaines soit pour améliorer l'échantillonnage en fonction de la base de sondage existante, soit comme liste pour réduire le nombre d'étapes dans le plan d'échantillonnage est un domaine important mis au premier plan pour faire l'objet de recherches. Avec les numéros de téléphone qui figurent sur le RA, il serait possible de réaliser plus d'entrevues téléphoniques.

L'utilisation du RA comme base de sondage pour d'autres enquêtes de Statistique Canada sera examinée. De plus, puisqu'on utilise actuellement les fichiers des sociétés de téléphone comme principale source de renseignements pour produire le RA, on dispose déjà de ces fichiers pour exploitation ultérieure. Le programme des enquêtes spéciales, l'enquête sociale générale et l'enquête existante sur la population active sont des enquêtes qui utilisent des fichiers des sociétés de téléphone ou qui en ont besoin.

Une autre application possible du RA à Statistique Canada consisterait à l'employer comme base de données sur le logement si on lui ajoutait les données sur le logement tirées du recensement de 1991 et les données provenant des rôles d'évaluation des municipalités, par exemple. L'existence d'une telle base de données pourrait réduire la quantité de renseignements sur le logement que l'on devrait recueillir dans les recensements à venir. Les besoins en données et la disponibilité de ces dernières doivent être étudiés.

6.6 Utilisations du registre des adresses à l'extérieur de Statistique Canada

Si le RA doit être utilisé à l'extérieur de Statistique Canada, il faut s'attaquer aux questions portant sur la confidentialité des fichiers sources et sur la possibilité de diffuser le RA, les conditions imposées doivent respecter les exigences de la Loi sur la statistique. Certains fichiers

pour certifier les totaux de logements pour le traitement ou pour des analyses de la migration. On étudiera si le RA devrait être utilisé avant ou après le jour du recensement et comment il pourrait être employé dans le cas des adresses pour lesquelles on ne peut établir des éléments géographiques qu'à un niveau supérieur au SD.

6.2 Rapports entre la Division de la géographie et le registre des adresses

Comme cela est manifeste dans la description de la méthodologie, lors de la création du RA on s'est basé beaucoup sur de nombreux produits de la Division de la géographie (p. ex., le Fichier principal de région, le Fichier de conversion des codes postaux). Nous examinerons la contribution de ces produits à l'établissement du RA ainsi que leurs limitations. On étudiera, pour tous les nouveaux produits élaborés par la Division de la géographie, l'utilisation possible de ces produits dans le RA en vue d'incorporer les besoins du RA directement dans les nouveaux produits. De plus, le RA sera intégré dans le SIG de la Division de la géographie.

Il se peut que le RA puisse fournir des indicateurs de mise à jour pour le FPR ou pour la délimitation des secteurs de dénombrement. Le RA pourrait être utilisé pour établir des ordres de priorité, particulièrement dans les régions à forte croissance ou dans les régions où les gammes de numéros de voirie dans les FPR sont de qualité médiocre. On pourrait utiliser des combinaisons de code postal/secteur de dénombrement ou de code postal/côté d'ilot provenant du RA pour effectuer la mise à jour du Fichier de conversion des codes postaux. Après chaque recensement, tous les ménages du recensement sont codés avec des centroïdes de côté d'ilot. Puisque la majeure partie des enregistrements du RA ont déjà été géocodés avant le recensement, le fait de coupler le RA avec le numéro de ménage du recensement réduira l'importance du travail de géocodage à effectuer manuellement après le recensement. Ce dernier projet est déjà en cours de réalisation.

6.3 Évaluation et amélioration des procédures et production de la documentation connexe

On est en train de préparer, pour le travail effectué jusqu'ici, un guide de l'utilisateur où l'on décrit les procédures et un guide technique qui renferme de la documentation relative aux programmes, à des problèmes types ainsi qu'à leurs solutions et à l'assurance de la qualité. Comme dans le cas de tout nouveau projet, beaucoup de choses sont apprises pendant le processus de création et les procédures sont élaborées au besoin et en fonction du temps ainsi que du budget disponibles. Après le fait, on peut habituellement réaliser des gains d'efficacité en examinant ces procédures.

Pour les procédures automatisées, les projets déjà en cours comprennent une utilisation plus efficace de ORACLE ou le choix d'un autre système, l'emploi d'ordinateurs de bureau plutôt que du gros ordinateur de Statistique Canada, la normalisation du filtre, des améliorations au PAAS, la fusion des unités de travail en bases de données provinciales, l'élimination de certains champs plus tôt dans le traitement, l'étude d'autres logiciels pour déterminer les codes postaux, l'amélioration de l'appariement d'une adresse avec un nom de lieu et l'amélioration du couplage entre le Fichier principal de région et les adresses en français.

Pour les procédures manuelles, on continuera d'améliorer le traitement des secteurs de dénombrement adjacents de chaque côté des limites de circonscriptions électorales fédérales et les cas où des numéros de voirie manquent sur les cartes produites par CAO. Le système de contrôle pour corriger les adresses sera aussi examiné afin de l'améliorer si cela est possible. Les numéros de téléphone ont été ajoutés à une étape ultérieure de la production du RA. Une évaluation complète de la couverture qu'ils offrent et de leur exactitude sera entreprise particulièrement en vue des utilisations possibles des numéros de téléphone dans le cadre du recensement et d'autres enquêtes de Statistique Canada. Pour ces dernières, l'accent initial sera placé sur la réalisation d'essais dans le contexte du remaniement imminent de l'enquête sur la population active.

4.10 Impression et production des cahiers (étape 14)

La dernière étape de la production était l'impression et l'assemblage des cahiers (étape 14) pour les près de 23,000 secteurs de dénombrement renfermant, à ce moment, 6,6 millions d'adresses. Des questions importantes qui ont été abordées comprenaient la rapidité et la qualité de l'impression (nous avons utilisé une imprimante à papier en continu), la durabilité des cahiers (ils comprenaient des couvertures avant et arrière et étaient agrafés) et les frais de compilation (les cahiers ont été assemblés et agrafés dans nos locaux).

5. ÉVALUATION POSTCENSITAIRE

L'évaluation postcensitaire peut être classée, en quatre domaines d'étude: opérations sur le terrain, saisie des données sur les cahiers du RA, mise à jour du RA et détermination de l'apport du RA aux améliorations de la couverture.

L'évaluation des opérations sur le terrain se concentrera sur l'efficacité de la formation, sur jusqu'à quel point le travail de conciliation a été complet et sur les causes des erreurs en vue d'améliorer la méthodologie pour les recensements à venir.

La saisie des données donnera deux résultats distincts. Premièrement, les adresses imprimées dans les cahiers seront supprimées si elles sont erronées et si elles sont valides, leur numéro de ménage du recensement sera saisi. Deuxièmement, les nouvelles adresses ajoutées par les recenseurs seront saisies. Il sera alors possible de calculer les taux de surdénombrement et de sous-dénombrement du RA ainsi que la contribution du RA à la couverture du recensement. On pourra examiner le cas des adresses classées dans le mauvais SD et remonter à la source de l'erreur. Le numéro de ménage du recensement nous permet d'étudier le nombre de personnes ajoutées ainsi que les caractéristiques des logements et des personnes.

Du point de vue coût, on calculera le coût unitaire par logement ajouté à cause de l'utilisation du RA en vue de déterminer le coût de création du RA et de son utilisation dans le cadre du recensement.

6. ORIENTATIONS FUTURES

Le registre des adresses (RA), bien qu'il ait été conçu à l'origine comme une des procédures visant à réduire le sous-dénombrement du recensement, est un projet évolutif avec un effet possible sur d'autres programmes de Statistique Canada ainsi que sur ceux d'autres organismes gouvernementaux.

Voici les objectifs plus immédiats pour le développement futur du RA: incorporer les adresses relevées pendant le dénombrement du recensement; évaluer l'efficacité du RA pour ce qui est de l'amélioration de la couverture du recensement de 1991; consigner par écrit et évaluer les activités de production et élaborer un plan à plus long terme pour le RA s'attaquant à son rapport coût-efficacité comme base de sondage de ménages, à la stratégie optimale de mise à jour et aux possibilités qu'offre son emploi par des organismes de l'extérieur du Bureau.

Compte tenu de ces lignes directrices, on a préparé un plan de projet qui est présenté ci-après sous six sujets principaux.

6.1 Rapports entre le recensement et le registre des adresses

En plus des possibilités qu'il offre pour améliorer la couverture, nous étudierons d'autres façons qui permettraient de mettre le RA à contribution pour le recensement. Certaines idées préliminaires à ce sujet comprennent la possibilité d'utiliser le RA comme fichier de contrôle du traitement, pour obtenir les numéros de téléphone qui seront employés lors d'un suivi éventuel, pour créer des chiffres de contrôle des logements dans un secteur de dénombrement,

Au cours de l'étape 12, nous traitions maintenant les adresses restantes qui n'avaient pu être apparées avec un seul SD, mais qui pouvaient l'être avec au moins deux SD (les SD candidats à un appartèlement) au cours de l'étape 8. Un ensemble complet de cartes conçues par cartographie assistée par ordinateur (CAO) a été produit pour le projet du RA. Au cours de l'étape Manuel-2, on examinait ces cartes pour trouver ces SD candidats afin d'affecter ces adresses restantes au SD approprié, dans la mesure du possible.

En général, le rapport entre l'appartèlement automatisé et l'appartèlement manuel était de 91% à 90%. La partie automatisée se répartissait de la façon suivante: 87% provenant du couplage RA/FPR au côté d'ilot et 40% du couplage RA/FCCP au SD. Pour le traitement manuel, la répartition des appartéments était la suivante: 30% provenant du couplage avec le côté d'ilot au cours de l'opération Manuel-1 et 60% du couplage avec le SD au cours de l'opération Manuel-2.

Bien que ORACLE ait constitué un véhicule approprié pour l'essai de 1989, son utilisation s'est révélée coûteuse et il a éventuellement constitué un goulot d'étranglement quand la production du RA battait son plein alors que ce dernier n'était qu'un des utilisateurs de cette base de données employée par tous les services du Bureau. ORACLE ne permettait d'avoir que de 8 à 10% des unités de travail accessibles en direct à un moment quelconque et le système devait exporter et importer continuellement des données sur les unités de travail afin de libérer de l'espace pour continuer le traitement. Nous avons donc constitué une seconde base de données ORACLE à l'usage exclusif de l'équipe travaillant au RA. En toute justice pour ORACLE, ce n'est pas tout le traitement à effectuer qui donnait des conditions propices à quelque système de gestion de base de données que ce soit. La création du produit était en cours et, par conséquent, de grandes parties des tables étaient examinées pour apporter des modifications radicales à des champs, pour éliminer des enregistrements en double et pour choisir des enregistrements à imprimer. ORACLE offrait une flexibilité considérable pour changer les procédures logicielles rapidement et pour en produire de nouvelles à mesure que l'étape de production se déroulait.

4.9 Utilisation du système de cartographie assistée par ordinateur (étape 13)

Le système de CAO était une nouvelle initiative de recherche pour le recensement de 1991 dont l'élaboration s'est faite concurremment avec celle du RA. Le système a produit toutes les cartes de secteurs de dénivellation dans les régions pour lesquelles il existe des FPR. Cela constituait un changement considérable par rapport au processus manuel de production de cartes utilisé auparavant. La CAO a aussi fourni, pour les SD, une structure dans laquelle les côtés d'ilot étaient situés dans les ilots et les ilots classés dans le SD (étape 13). Aux fins du RA, on a préparé un programme qui dérive de la CAO pour classer les logements dans un côté d'ilot. Cette opération était nécessaire pour structurer les listes d'adresses afin qu'elles correspondent plus étroitement à la façon dont les recenseurs dressent leurs listes.

Le système de CAO était entièrement mis en place au moment de la production du RA. Afin d'assurer la compatibilité avec ce logiciel, nous avons utilisé la version du FPR qui était employée par le système de CAO. Toutefois, aucune donnée sur la structure n'était affectée à une petite portion des côtés d'ilot. Pour tout SD où ce pourcentage était supérieur à 5%, nous reprisons le traitement effectué par le système de CAO pour cette unité de travail, si le temps le permettait, ou nous utilisons un autre système, celui qui emploie l'algorithme d'affectation des points dans un polygone (Point in Polygon Assignments (PIPA)), qui situe les côtés d'ilot dans leur SD. Bien que le système PIPA faisait passer les adresses de la partie structurée du cahier du RA (basée sur le codage des côtés d'ilot) à la partie non structurée (basée sur le codage des SD), les adresses posant des problèmes lors du processus de sélection pour l'impression pouvaient au moins être conservées, ce qui n'était pas le cas quand les données sur l'ordonnancement manquaient.

À nouveau, c'est Brampton qui a été utilisé pour l'essai. L'analyse de l'appariement code postal/SD a permis d'établir que 38% des codes postaux pouvaient être attribués, de façon unique, à un SD de 1991. Le couplage de ces codes postaux à des enregistrements du RA qui n'étaient pas apparés à un côté d'îlot a permis d'augmenter d'un autre 5% le nombre total d'appariements. Dans l'ensemble, le taux des appariements automatisés a augmenté pour atteindre 89% (84% avec le côté d'îlot et 5% avec le SD), en hausse par rapport à 64% lors de l'essai de septembre 1989, ce qui réduit de près de la moitié l'importance des opérations manuelles à effectuer.

4.7 Chargement dans la base (étape 9)

Lors de l'essai de 1989, pour faciliter les consultations et en prévision d'une utilisation ultérieure, nous avons utilisé ORACLE comme système de gestion de base de données et nous l'avons employé à nouveau pour la production du RA en 1991. L'étape du chargement des données dans ORACLE (étape 9) comprenait la transformation du fichier jusqu'ici séquentiel en quatre fichiers de composantes distincts, un pour chacun des éléments suivants: municipalité, côté d'îlot, rue et adresse.

4.8 Travaux de bureau (étapes 10, 11 et 12)

Lors de l'essai de 1989, le travail de bureau consistait à examiner toutes les combinaisons uniques de nom de rue/d'indicateur de genre de rue/de sens de rue tirés des enregistrements tant du FPR que du RA avec un relevé du nombre d'enregistrements dans le RA pour chaque combinaison de rues. L'objectif de ce travail était de remplacer une combinaison de rues du RA pour laquelle il n'y avait pas d'appariement par la combinaison valable du FPR. La comparaison de combinaisons de rues semblables et la détermination de celles qui auraient, en fait, dû être identiques, permettrait d'apparier manuellement des enregistrements du RA non codés jusqu'à ce moment avec un côté d'îlot particulier. Cette procédure avait donné de bons résultats en 1989 et s'était révélée utile pour régler deux situations qui posaient des difficultés: les cas où il y avait des écarts considérables dans l'orthographe du nom de la rue et ceux où le champ du nom de la rue dans le RA contenait à la fois le nom de la rue et une forme abrégée d'indicateur de genre de rue que le logiciel PAAS n'avait pas compris lors de l'analyse de l'adresse.

Nous avons accru la portée de ce travail de bureau (étape 10) afin de comparer des combinaisons de rues du RA avec d'autres combinaisons de rues du RA semblables pour traiter les cas où il se pourrait que l'on trouve un certain nombre de variations orthographiques du nom d'une rue particulière dans le RA sans équivalents dans le FPR. Cet accroissement des possibilités a permis de coder plus d'adresses au côté d'îlot.

Pour résumer, au cours de ce premier travail de bureau (Manuel-1), toutes les adresses qui n'ont pas été codées automatiquement au côté d'îlot dans l'étape 7 (c'est-à-dire, les adresses codées automatiquement au SD dans l'étape 8 et celles qui ne sont pas encore codées) ont été examinées en vue d'un codage manuel possible au côté d'îlot.

Suite au premier travail de bureau (Manuel-1), nous avons ajouté une étape de compression (étape 11), qui a été appliquée à tous les enregistrements codés au côté d'îlot. Pour chaque valeur unique de nom de rue/indicateur de genre de rue/sens de rue pour une unité de travail, nous avons vérifié tous les enregistrements d'adresses correspondants afin de nous assurer qu'ils sont uniques en utilisant le numéro d'appariement comme clé. Quand des enregistrements multiples étaient rencontrés, ils étaient regroupés avec toutes les données pertinentes combinées en un seul enregistrement, une étape additionnelle d'élimination d'enregistrements en double.

En conséquence, à la fin de l'étape 10, la base de données renfermait des adresses codées automatiquement ou manuellement au côté d'îlot, codées automatiquement au SD ou pas encore codées.

était composée du nom de la rue/de la région de tri d'acheminement (RTA)/du drapeau de numéros de voirie impairs ou pairs. La deuxième poche comprenait le code postal/le drapeau de numéros de voirie impairs ou pairs, ce qui permettait d'apparier d'après le code postal les adresses mal analysées. La troisième poche comprenait la version NYSSIS du nom de la rue/du drapeau de numéros de voirie impairs ou pairs, ce qui permettait de considérer comme des appar- tiements éventuels des enregistrements renfermant des variations orthographiques du nom de la rue ou qui n'ont pas de code postal.

Les règles de fonction établies dans le cas des appartements partiels pour le nom de la rue, pour le nom de la municipalité et pour les trois derniers caractères du code postal ont été tirées directement de notre application CANLINK existante utilisée pour l'élimination interne des données en double, où elles avaient déjà démontré leur efficacité.

Cependant, au cours de l'étape de la production du RA, nous avons éprouvé de la difficulté à apparier les enregistrements contenus dans trois FPR: Red Deer, St. Thomas et Charny. Dans les trois cas, c'est le manque de données sur les numéros de voirie dans le FPR qui a créé les difficultés. Sachant qu'il faudrait effectuer beaucoup de travail de bureau pour éliminer ces difficultés, une opération sur le terrain a été lancée en décembre 1990 afin de mettre à jour les cartes produites à l'aide du système de cartographie assistée par ordinateur (CAO). La Division de la géographie a envoyé les cartes produites par CAO aux bureaux régionaux où des employés ont ajouté les gammes de numéros de voirie manquants. Ces cartes mises à jour ont ensuite été retournées à la Division de la géographie pour être incluses dans la prochaine série de mises à jour des FPR. Pour la création du RA, les gammes de numéros de voirie pour les trois FPR ont été utilisées dans des opérations manuelles effectuées par des employés de bureau.

Au niveau de l'appariement, le succès était fort semblable dans toutes les provinces sauf le Québec. Au Québec, l'appariement automatique au côté d'ilot a diminué d'environ 10 à 12% pour atteindre 73%, puisque cette opération n'était pas aussi efficace lors du traitement des adresses en français qu'elle ne l'était dans le cas des adresses en anglais. On a trouvé trois situa- tions qui causaient la diminution du taux d'appariement automatique: l'utilisation ou la non utilisation d'articles dans les noms de rue (p. ex., Savane, de la Savane, la Savane), l'utilisation d'un nom complet de personne comme nom de rue avec beaucoup de variantes orthographiques (p. ex., Jean-François Bélanger, J. F. Bélanger, Jean F. Bélanger) et le manque d'indicateurs de genre de rue. C'est pourquoi, les opérations de bureau décrites ci-dessous, particulièrement la première, avaient une importance accrue pour l'appariement au Québec par rapport à la situation dans les autres provinces.

Lors du traitement du RA/FPR à l'aide du logiciel CANLINK, il n'y a eu qu'une difficulté soit le fait que le nombre maximum d'enregistrements permis dans une poche interne à CANLINK pouvait être dépassé. La solution adoptée consistait à déterminer les rues qui étaient la source de la difficulté à partir du rapport sur la poche (il s'agissait toujours d'artères princi- pales) et à préparer des programmes spéciaux de pré-traitement qui ajouteraient le cinquième caractère du code postal lors du calcul de la valeur de la poche pour ces rues afin qu'une distinction puisse être établie plus facilement. Cela avait pour effet de réduire le nombre d'enregistrements dans la poche.

4.6 Couplage RA/FCCP (étape 8)

Dans cette étape (étape 8) on tentait d'obtenir un couplage automatisé avec le secteur de dénombrement (SD) approprié pour les adresses qui n'avaient pu être apparées, à l'aide du FPR, aux côtes d'ilot pendant l'étape 7.

Les principales données en entrée étaient le Fichier de conversion des codes postaux (FCCP), qui donnait la correspondance entre les codes postaux et les SD de 1986 et le fichier de corres- pondance entre SD de 1986 et de 1991. En faisant l'appariement de ces deux fichiers nous pouvions déterminer les codes postaux qui ne correspondaient qu'à un seul SD de 1991, ainsi que les codes postaux qui correspondaient à au moins deux SD possibles pour 1991 et dont le cas allait être réglé, par des opérations manuelles, au cours de l'étape 12.

à des fins d'appariement et des poids ont été attribués quand il y a accord ou désaccord pour chaque élément. L'élaboration de niveaux d'accord partiel pour le nom de la rue, pour le nom de la municipalité et pour les trois derniers caractères du code postal tenait compte des variations orthographiques et des transpositions de lettres dans les champs. L'étape CANLINK a entraîné une autre réduction du nombre d'enregistrements, il n'y en avait que 6,7 millions après cette étape. On trouve plus de détails sur l'utilisation de CANLINK pour l'élimination des adresses en double dans Drew et coll. (1988), où l'on décrit l'application de ce logiciel dans le cadre de l'essai de novembre 1987.

4.5 Couplage RA/FPR (étape 7)

La stratégie utilisée pour coupler les adresses à leur côté d'ilot respectif constituait la principale préoccupation après l'essai de 1989. À cause de la diminution de 11% dans la couverture qui est passée de 84% à 73% comparativement à l'essai de 1987, il fallait effectuer une étude complète et peut-être employer une nouvelle approche. L'autre préoccupation en matière de procédure était le fait que l'appariement automatisé n'avait permis d'obtenir que 80% des enregistrements appariés alors que le reste (20%) de l'appariement était attribuable à du travail de bureau. Cette situation aurait représenté une lourde charge de travail manuel quand la production du RA battait son plein. Afin de surmonter ces deux préoccupations, on a élaboré une autre application de CANLINK pour le couplage RA/FPR (étape 7). Les fichiers originaux utilisés dans le cadre de l'essai de 1989 pour Brampton existaient encore, c'est donc sur cette localité qu'a porté l'essai visant à élaborer cette étape. La méthode révisée a donné 10% de plus d'appariements, ce qui a ramené la couverture aux niveaux de 1987. De plus, l'appariement automatisé permettrait de réaliser 97% des appariements, 3% étant effectués par des employés de bureau, une amélioration importante par rapport à la répartition antérieure de 80%-20%. Suite à ces résultats, la méthode CANLINK a été adoptée pour la réalisation des travaux pour le recensement.

Lors de l'élaboration de la nouvelle stratégie d'appariement, le premier domaine d'étude comprenait une comparaison du contenu des champs qui seraient utilisés pour effectuer l'appariement. Cette opération a révélé certaines anomalies qui pouvaient être corrigées avant l'utilisation des fichiers afin d'améliorer le nombre de couplages. Les modifications du traitement des champs existants portaient sur les domaines suivants: suppression des espaces entre les noms de rue composés; alignement des sens des rues et des numéros de voirie; conversion des noms de rue numériques en chiffres (dans le FPR); suppression des caractères spéciaux dans les noms de rue (dans le FPR); correction des variantes orthographiques dans les noms de municipalité (dans le RA) et une reconstitution de certaines traductions effectuées par le PAAS pour des noms de rue (dans le RA). Plusieurs nouveaux champs ont aussi été créés: des versions NYSIS (New York State Identification and Intelligence System) et SOUNDEX du nom de rue, à l'aide de deux logiciels de codage phonétique utilisés pour éliminer les effets des erreurs d'orthographe courantes (Statistique Canada 1989d); un drapeau de nom de rue en double (dans le FPR) pour signaler les situations où un nom de rue n'est pas unique; un drapeau de rue unidirectionnelle (dans le FPR) pour signaler les rues pour lesquelles un seul sens de rue avait été codé et un drapeau de nom de rue officiel (dans le RA) pour signaler que le nom de la rue correspondait à un nom officiel de rue dans le FPR. Les enregistrements du FPR ne contenaient que des données sur les rues, nous leur avons donc ajouté le nom de la subdivision de recensement ainsi qu'un code de province puis nous avons tenté d'attribuer des codes postaux aux numéros de voirie dans des côtes d'ilot. Quand les codes postaux pour les numéros de voirie "de départ" et "d'arrivée" différaient, nous avons produit des sous-côtes d'ilot pour chaque code postal unique.

Pour cette application, trois poches distinctes ont été créées pour chaque enregistrement produisant, effectivement, trois exemplaires des fichiers. La poche primaire était celle à laquelle s'appliquaient les conditions les plus rigoureuses et elle était conçue pour trouver rapidement toutes les possibilités d'un bon appariement lors du premier passage des fichiers. Cette poche

éliminer, l'un après l'autre, par un filtrage additionnel jusqu'à ce qu'un taux d'erreur inférieur à 5% soit atteint. Comme tout enregistrement d'adresse qui était rejeté lors de la normalisation des adresses était éliminé de tout traitement ultérieur, il était essentiel que le taux de réussite du PAAS soit le plus élevé possible.

L'étape PCODE (étape 3) utilisait le progiciel du système automatisé d'établissement des codes postaux (PCODE) (Statistique Canada 1989a) pour confirmer et pour produire les codes postaux. Ce logiciel n'était pas tout à fait aussi efficace que le logiciel PAAS pour analyser les adresses et il ne pouvait confirmer ou ajouter des codes postaux que pour 84% des sorties du PAAS. Ce progiciel a confirmé 78% des codes postaux et il en a modifié un autre 6%. Seulement 0,003% des enregistrements administratifs d'origine avaient été fournis sans code postal. Il était essentiel de disposer des codes postaux exacts parce que ces derniers allaient être utilisés pour le choix de l'unité de travail au cours de l'étape suivante.

Deux problèmes ont surgi lors de l'étape PCODE au moment de la production du RA. S'il n'y avait pas d'élément municipalité ou province pour une adresse, le logiciel continuait de tenter de trouver un code postal plutôt que de suspendre le traitement. Par conséquent, des temps de traitement considérables pouvaient être passés à essayer de trouver des codes postaux. Pour résoudre ce problème, nous avons inclus dans l'opération de FILTRAGE une étape pour ajouter le nom de la municipalité et celui de la province. Le deuxième problème se produisait quand un nom de rue était numérique, car alors le temps de traitement par adresse quadruplait. Ce problème n'a pu être résolu et pour l'éliminer il faudra modifier le logiciel PCODE.

4.3 Choix de l'unité de travail (étape 4)

Au cours de cette étape le pays a été divisé, au moyen des codes postaux, en unités de travail administrables aux fins du traitement, la taille des unités de travail étant basée sur l'efficacité du logiciel CANLINK pour coupler de nombreux gros fichiers. On a adopté une division géographique par unité de travail telle que ces derniers comprenaient entre 100,000 et 150,000 logements d'après les données du recensement de 1986. Les unités de travail étaient formées à partir d'un seul FPR (pour une ville de taille moyenne), à partir d'ensembles de FPR adjacents (pour de petites villes/cantons), ou à partir de parties d'un FPR (pour une grande ville). Le Fichier de conversion des codes postaux (FCCP) de la Division de la géographie qui couple les codes postaux avec des éléments détaillés de géographie du recensement a été utilisé pour effectuer ce partage lors de l'étape de SÉLECTION (étape 4). Une fois le partage terminé, il y avait 105 unités de travail distincts et le nombre d'adresses originales avait été ramené de 43,4 millions à 20,5 millions, les adresses éliminées ayant des codes postaux à l'extérieur des régions couvertes par les FPR (c.-à-d. dans les plus petites villes et dans les régions rurales).

4.4 Élimination des adresses en double (étapes 5 et 6)

À fin de supprimer les adresses qui figuraient plus d'une fois dans les fichiers sources, une opération d'élimination d'adresses en double a été effectuée en deux étapes; un appariement exact avec DEEXACT (étape 5) et un appariement probabiliste à l'aide du logiciel CANLINK (étape 6).

Au cours de l'étape DEEXACT on utilisait la clé de recherche d'adresses (CRA) produite par le logiciel PAAS et tous les enregistrements avec une CRA identique ont été comprimés en un seul enregistrement. Avec DEEXACT, les 20,5 millions d'enregistrements dont on disposait après l'étape de SÉLECTION ont été ramenés à 10,1 millions d'enregistrements. Cette réduction fait ressortir l'importance d'effectuer la normalisation des adresses. Dans l'étape 6 on utilisait le logiciel général de couplage d'enregistrements CANLINK (Statistique Canada 1989b). Ce logiciel réunit des enregistrements "proches" en groupes appelés "poches" et seuls les enregistrements dans la même poche sont effectivement appariés. Pour cette application, c'est le numéro de voirie qui a été utilisé comme poche. Les éléments constitutants de l'adresse (nom de la rue, nom de la municipalité, code postal, etc.) ont été utilisés

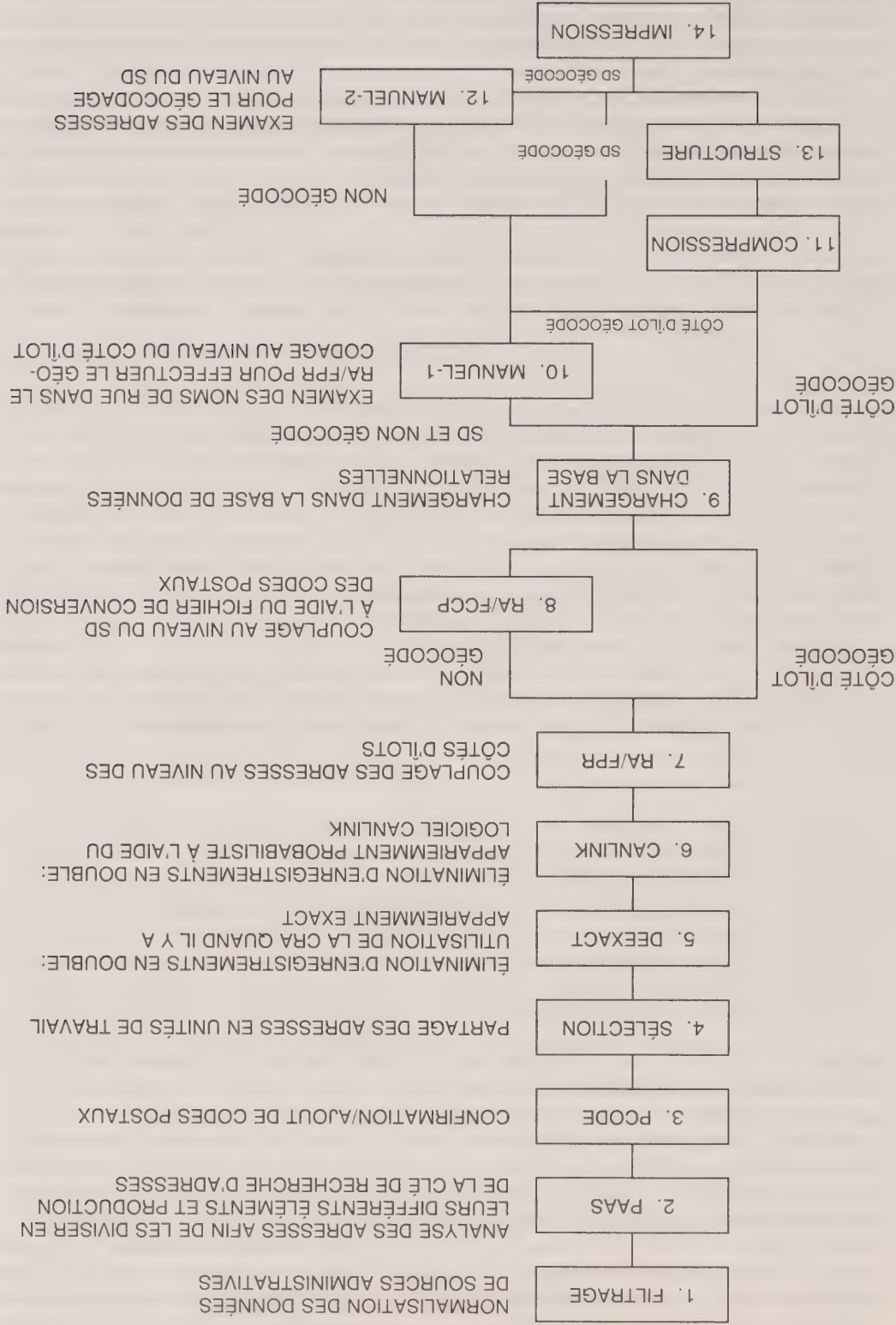


Figure 2. Aperçu de la méthodologie.

ou pour toute autre raison. Pour toutes les adresses valables le recenseur inscrivait le numéro de ménage du recensement dans le cahier. Un numéro de téléphone correspondant à l'adresse, s'il était disponible, était imprimé à l'avance dans la dernière colonne du cahier afin d'aider le recenseur dans toute opération nécessaire pour effectuer le suivi du recensement.

4. MÉTHODOLOGIE

Dans la présente section, nous décrivons la création du RA. La figure 2 donne un aperçu des étapes que comprend cette opération.

4.1 Aperçu de la méthodologie

On a tout d'abord effectué la normalisation en éléments constituant ordonnés des adresses à structure non imposée contenues dans les fichiers sources (étapes 1 et 2) en vue de l'utilisation du logiciel qui effectue les étapes ultérieures. Puis, les codes postaux ont été confirmés ou corrigés (étape 3) afin que les régions ou unités de travail pour lesquels le RA allait être créé puissent être choisis à partir de toutes les adresses et emplacements contenus dans les fichiers sources (étape 4). Parce que la même adresse pouvait être contenue dans plus d'un fichier ou plus d'une fois dans le même fichier, on a effectué la suppression des adresses en double en se basant sur des appariements tant exacts que probabilistes (étapes 5 et 6). Ensuite, on a procédé au couplage automatisé des adresses au niveau des côtes d'îlots à l'aide du Fichier principal de région (étape 7) ou, quand cela n'était pas possible, au secteur de dénombrement (SD) à l'aide du Fichier de conversion des codes postaux (étape 8). Après avoir chargé les adresses dans un système de gestion de base de données (étape 9), des coupplages manuels ont été effectués entre les adresses et les côtes d'îlot (étapes 10 et 11) ou entre les adresses et les SD (étape 12). Les adresses dans chaque SD ont ensuite été classées par côté d'îlot et à l'intérieur des côtes d'îlot (étape 13) avant d'être imprimées et assemblées en cahiers par SD (étape 14) pour utilisation dans le cadre du recensement.

4.2 Normalisation des adresses (étapes 1, 2 et 3)

Le système d'analyse des adresses postales (PAAS – étape 2 de la figure 2) (Statistique Canada 1989c) remplissait deux tâches: tout d'abord il séparait les adresses à structure non imposée provenant des fichiers sources en leurs éléments constitutants (nom de rue, numéro de voirie, genre de rue, sens de la rue, numéro d'appartement, municipalité, province, code postal) et composait la clé de recherche d'adresses (CRA). La CRA est un enchaînement ordonné de tous les éléments qui composent une adresse et elle est utilisée pendant les opérations visant à éliminer les adresses en double.

Bien que le PAAS fut un excellent produit, l'analyse des résultats du prototype de 1989 avait révélé certains défauts qui, selon nous, pourraient être réglés en traitant le contenu des fichiers administratifs avant d'utiliser le logiciel général. Cette étape de FILTRAGE (étape 1) portait sur les tâches suivantes: élimination des caractères spéciaux que le PAAS refusait de traiter, regroupement des éléments de l'adresse afin de la rendre compatible avec le PAAS, traduction des indications abrégées de genre de rue en genres acceptables, introduction de virgules entre les éléments de la rue et de la municipalité des adresses à structure non imposée afin que le PAAS puisse mieux les comprendre, élimination des zéros en tête dans les numéros de voirie et les noms d'adresses numériques et ajout du nom de la municipalité et de la province.

Les étapes de FILTRAGE et PAAS ont été appliquées de façon itérative. La première étape consistait à découvrir quelles anomalies devaient être filtrées pour chaque source de données administratives. Si le taux d'erreur du PAAS après le filtrage était supérieur à 5%, les enregistrements erronés étaient examinés afin de trouver les problèmes qui se répétaient en vue de les

3.3 Cahiers du registre des adresses

Le produit final était un ensemble de cahiers d'adresses résidentielles, un pour chaque secteur de dénombrement, qui englobait toutes les régions urbaines du Canada pour lesquelles un fichier principal de région existait. La figure 1 renferme un exemple fictif (en format réduit) d'une page d'un cahier du RA.

Chaque cahier était divisé en deux sections: une partie structurée et une partie non structurée. La partie structurée renfermait toutes les adresses liées à un côté d'ilot et tous les côtés d'ilot y étaient classés en îlots dans le SD. Le classement correspondait aux renseignements qui figuraient sur la carte utilisée par le recenseur pour dresser la liste du SD dans son Registre des visites (RV). La partie non structurée contenait les adresses qui ne pouvaient être liées qu'au SD plutôt qu'à un côté d'ilot. Ces adresses étaient classées par numéros de voirie impairs/pairs pour un même nom de rue. Le nombre d'adresses était réparti dans la proportion suivante: 90%-10% entre les données structurées et les données non structurées.

En plus des données sur les adresses, chaque page d'un cahier de RA comprenait une série de colonnes à utiliser lors de l'opération de conciliation entre le RA et le RV. Au cours de la conciliation, le recenseur comparait manuellement le Registre des visites au RA afin de déterminer où il y avait correspondance et où il n'y en avait pas. Si l'adresse n'était que dans le RV, elle était ajoutée au RA (sous-dénombrement dans le RA). Si l'adresse ne figurait que sur le RA, le recenseur devait habituellement régler le problème sur le terrain. Cette adresse était donc désignée soit comme une nouvelle adresse que le recenseur devait dénombrer dans le cadre du recensement (sous-dénombrement au niveau du recensement), soit comme une adresse invalide classée selon le genre d'erreur (surdénombrement dans le RA). Les adresses étaient considérées invalides s'il s'agissait d'adresses en double, si elles se trouvaient à l'extérieur du SD

REGISTRE DES ADRESSES Protégé PROVINCE CEF 038 NV 0 SD 261 Page 21 de 22

N° d'ot		Adresse		N° de ménage		N° de ménage à la livraison		Suivi		Non valide		N° de réf.		Numéro de téléphone	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	N° de voiture	Rue	N° d'app.	N° de ménage	N° de ménage à la livraison	place requis	En double	En dehors du SD	Autre	N° du RA					

4	23	PRINCIPALE	RU	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
4	19	PRINCIPALE	RU	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
4	15	PRINCIPALE	RU	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
4	11	PRINCIPALE	RU	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
4	7	PRINCIPALE	RU	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	30	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	34	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	60	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	64	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	68	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	72	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	76	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	80	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	84	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	88	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	92	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	96	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	100	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	108	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	112	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	116	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	120	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588
5	124	CENTRE	BV	1044566	5551111	1044564	5561234	1044562	5552321	1044559	1044583	7475739	5552222	1044581	5556942	1019615	1019617	1019618	5564261	1019627	1019629	7478765	5556942	1019636	1019640	1019642	7476789	5568765	5559999	7473456	1019579	1019581	5557171	5558888	1019581	7462009	1019586	7450235	5569630	1019588

3. SOURCES ET PRODUIT

La production du registre des adresses (RA) a commencé en avril 1990 et s'est terminée avec l'agrégage du dernier cahier à la mi-mai 1991, quand 22,756 cahiers renfermant 6.6 millions d'adresses avaient été compilés pour utilisation au cours des opérations de collecte des données du recensement.

3.1 Sources administratives

Suite à l'essai de septembre 1989, on a conclu que, dans la mesure du possible, l'on devrait utiliser les quatre sources administratives mentionnées ci-après comme sources d'adresses pour créer le RA: fichiers de facturation des sociétés de téléphone, rôles d'évaluation des municipalités, fichiers de facturation des sociétés d'électricité et le fichier des déclarations d'impôt sur le revenu des particuliers (T1). Toutefois, ce n'est qu'en Nouvelle-Ecosse, au Nouveau-Brunswick et dans huit grands centres urbains de l'Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton et Windsor) que l'on pouvait utiliser ces quatre sources de données. À cause de la multiplicité des fichiers, de leur coût et des refus, seulement trois sources ont été utilisées pour Terre-Neuve, pour le Québec, pour le Manitoba, pour l'Alberta (fichiers des sociétés de téléphone, des sociétés d'électricité et de l'impôt) ainsi que pour Regina et pour le reste de l'Ontario (fichiers des sociétés de téléphone, rôles d'évaluation et fichier de l'impôt). Pour Saskatchewan, seuls les fichiers des sociétés de téléphone et de l'impôt sur le revenu étaient disponibles. Les principaux fichiers sources utilisés par le gouvernement de la Colombie-Britannique étaient ceux des sociétés de téléphone et des sociétés d'électricité, bien que l'on ait aussi employé les fichiers des immatriculations de véhicules, ceux des câblo-distributeurs et les listes électorales.

3.2 Sources de données géographiques

Lors de la création du RA, nous avons largement utilisé un système et trois fichiers de la Division de la géographie.

i. Le Fichier principal de région (FPR) (Statistique Canada 1988) est un réseau numérisé d'éléments (renfermant des rues, des voies ferrées, des fleuves et des rivières, etc.) pour de grandes et de moyennes régions urbaines, dont la population est généralement de 50,000 personnes ou plus. Pour le RA, nous étions intéressés par les données sur les rues qui comprenaient le nom de la rue et les gammes de numéros de voirie qui pouvaient être utilisés pour situer des adresses particulières sur un côté d'ilot, le principal élément pour effectuer le couplage.

ii. Le système de cartographie assistée par ordinateur (CAO) groupe les côtés d'ilot en ilots et les ilots pour former un secteur de dénombrement (SD) du recensement. Le système de CAO a été utilisé pour classer les adresses dans les cahiers du RA. Les recenseurs qui ont travaillé au recensement de 1991 ont utilisé les cartes de SD produites par le système de CAO. Dans le cas du RA, les cartes pour toutes les régions couvertes par un FPR ont été utilisées lors de la deuxième opération effectuée par les employés de bureau.

iii. Le fichier de conversion des codes postaux (FCCP) de 1990 (Statistique Canada 1991) est un fichier qui contient tous les codes postaux au pays, chacun étant couplé à un SD ou à une série de SD du recensement de 1986. Ces données en entrée ont été utilisées pour effectuer le couplage secondaire des adresses au niveau du SD.

iv. Le fichier de correspondance entre SD de 1986/1991 établit le lien entre les éléments géographiques des SD de 1986 et ceux de 1991. Ce fichier a été utilisé pour effectuer le couplage secondaire au niveau du SD et pour la deuxième opération effectuée par les employés de bureau.

(Royce et Drew 1988). On a estimé que 34,000 logements occupés et 68,000 personnes seraient ajoutés, à la suite de la production du RA, aux centres urbains, qui comptent une grande population ou une population moyenne, pour lesquels ce RA serait produit (ces centres urbains représentent les régions pour lesquelles le fichier principal de région existe, c.-à-d. qu'ils renferment environ 65% de la population canadienne). Cela représenterait une amélioration de la couverture de 0.26 point (le taux national de sous-dénombrement en 1986 étant estimé à 3.21%). Par rapport aux deux tentatives antérieures pour produire un RA, on a prouvé que les coûts, pour le recensement, étaient faibles à cause de la méthode très automatisée et de l'avantage démontré. De plus, le risque était minimisé puisque la méthode de collecte traditionnelle serait encore utilisée. Basé sur ce coût, sur l'avantage offert et sur l'évaluation du risque, on a approuvé la création d'un RA pour le recensement de 1991.

Deux questions se sont posées après l'essai de novembre 1987. Premièrement, le classement des adresses dans les cahiers du RA produits pour chaque secteur de dénombrement (SD) ne correspondait pas à leur ordre dans les Registres des visites, ce qui faisait de la conciliation une tâche ennuyeuse et prenant beaucoup de temps. Deuxièmement, le surdénombrement global qui s'établissait à 17% semblait encore trop élevé et il fallait plus d'efforts pour éliminer les enregistrements mal classés ou en double. On s'est attaqué à ces deux problèmes en améliorant les méthodes utilisées pour apparier le RA aux éléments géographiques du recensement. Plutôt que de coupler les adresses seulement aux SD comme on l'avait fait lors de l'essai de novembre, on a élaboré des procédures pour apparier le RA aux côtes d'ilot qui figurent dans le fichier principal de région (FPR) (Statistique Canada 1988). Un algorithme a été produit afin de trier les adresses par ilot et, à l'intérieur d'un ilot, dans l'ordre dans lequel le recenseur les rencontrerait s'il parcourait le SD à pied.

2.2 L'essai de septembre 1989 visant à améliorer les procédures

Un autre essai important a été réalisé en septembre 1989, il portait sur quatre villes de taille différente: Moncton, Laval, Brampton et Calgary. Chacune de ces villes a été choisie à cause des difficultés particulières qui pourraient s'y présenter d'après les renseignements obtenus lors de l'essai de novembre 1987. Les résultats (Dick 1990) ont montré une diminution importante dans la couverture qui est passée de 84% lors de l'essai de 1987 à 73%, un résultat encourageant. Par contre, cet essai a fait ressortir une réduction considérable du surdénombrement qui a diminué de 17% à 8%. Il est important de remarquer qu'en dépit de la couverture réduite du RA, le rendement de ce dernier comme outil pour améliorer la couverture du recensement était encore acceptable. Après analyse, on a trouvé que la nouvelle opération de géocodage était problématique, tant pour ce qui est des coûts élevés, puisqu'elle comportait beaucoup d'opérations effectuées par des employés de bureau, que pour sa qualité. Les étapes du géocodage ont donc été améliorées en vue de la production du RA. L'adoption du logiciel de couplage d'enregistrements CANLINK (Statistique Canada 1989b) afin d'améliorer la qualité et de réduire les coûts du couplage du RA avec le FPR a été un aspect clé de cette amélioration.

2.3 Accord avec la province de la Colombie-Britannique

Au Ministry of Finance and Corporate Relations de la Colombie-Britannique on s'inquiétait du taux élevé de sous-dénombrement dans la province lors du recensement de 1986 (4.49% en 1986, en hausse par rapport à 3.16% en 1981, pour l'ensemble de la population de la province) (Statistique Canada 1990). Statistique Canada a conclu à un accord avec la Planning and Statistics Division (l'organisme statistique provincial) du ministère afin d'aider à réduire le sous-dénombrement en Colombie-Britannique lors du recensement de 1991. Dans le cadre de ce contrat, le registre des adresses a été accru afin qu'il comprenne des centres urbains plus petits en Colombie-Britannique, augmentant ainsi la population couverte de 62% à 88%.

réalisé par envoi postal plutôt qu'avec la méthode traditionnelle de livraison. Cette étude a conclu que la nouvelle méthode de collecte des données du recensement serait moins dispendieuse seulement si la qualité du registre des adresses était telle qu'on ne devrait effectuer qu'un minimum de mises à jour sur le terrain avant le recensement. Deux petits registres-pilotes créés au début de 1987 ont permis d'établir la couverture du registre des adresses à 90-95%, ce qui était inacceptable sans mise à jour sur le terrain (Drew et coll. 1987), éliminant ainsi l'utilisation d'un registre des adresses pour un recensement par envoi postal.

Toutefois, les deux registres-pilotes ont permis d'établir les possibilités qu'offrirait un registre des adresses pour aider à améliorer la couverture quand il était employé avec la méthode traditionnelle de livraison. Cela cadrait bien avec l'apparition de l'amélioration de la couverture comme un des éléments prioritaires pour le recensement de 1991. Les résultats de la contre-vérification des dossiers pour le recensement de 1986 avaient montré une augmentation considérable dans le taux de sous-dénombrement comparativement aux recensements antérieurs (de 2.01% en 1981 à 3.21% en 1986 pour la population totale du pays; de 2.08% en 1981 à 3.28% en 1986 pour la population urbaine à l'échelle nationale) (Statistique Canada 1990). Il a donc été décidé que le projet de recherche devrait se concentrer sur l'élaboration du registre des adresses pour utilisation afin d'améliorer la couverture du recensement de 1991.

Dans la section ci-après on décrit les deux principaux essais effectués pour élaborer et améliorer les procédures utilisées afin de créer le registre des adresses pour le recensement de 1991. De plus, dans la deuxième section on décrit sommairement l'accord conclu avec la province de la Colombie-Britannique visant à augmenter le registre des adresses. Dans la troisième section on présente les sources administratives et géographiques utilisées lors du processus de production ainsi que la structure et le contenu des cahiers du registre des adresses, le produit final utilisé par les recenseurs sur le terrain. Dans la quatrième section on décrit la méthodologie employée pour exploiter les sources disponibles afin de produire les cahiers du registre des adresses. Dans la cinquième section on discute de l'évaluation postcensitaire proposée alors que dans la dernière section on présente les perspectives futures pour le registre des adresses. Un rapport distinct renfermant une évaluation détaillée de la méthodologie sera produit plus tard.

2. DONNÉES DE BASE

2.1 L'essai de novembre 1987 de méthodes d'amélioration de la couverture

Un essai important de l'utilisation du registre des adresses (RA) comme outil pour améliorer la couverture a été réalisé en novembre 1987 dans cinq grandes villes où l'on trouve un bureau régional. Cet essai avait été conçu pour estimer à la fois le sous-dénombrement et le surdénombrement des logements pour la méthode traditionnelle de listage du recensement et pour deux méthodes expérimentales utilisant un registre des adresses, afin d'améliorer la couverture du recensement: le post-listage et le pré-listage. Quand il appliquait la méthode du post-listage, le recenseur compilait la liste des logements selon les méthodes habituelles du recensement (en créant un Registre des visites) puis il la conciliait avec une liste de logements pour le secteur de dénombrement (SD) produite à partir du RA. Des suivis sur le terrain ont été effectués lorsqu'il y avait des écarts, au niveau des adresses, entre les deux listes. Pour la méthode du pré-listage, on a remis le RA au recenseur à l'avance et ce dernier en a effectué la mise à jour lors d'une prospection du SD afin de créer la liste finale des logements.

Les résultats (van Baaren 1988) concluaient que la méthode du post-listage était la méthode la plus efficace pour améliorer la couverture. Cette méthode appliquée comme simple ajout au processus normal de dénombrement du recensement était totalement sûre. Si pour une raison quelconque nous ne produisons pas le RA (soit en entier, soit en partie) à temps pour le recensement de 1991, l'étape de conciliation à l'aide du RA pouvait tout simplement être éliminée sans que cela ait un effet sur le processus de dénombrement traditionnel. Les données d'essai ont aussi fourni des estimations de l'importance de l'amélioration de la couverture et des coûts

La création d'un registre d'adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991

L. SWAIN, J.D. DREW, B. LAFRANCE et K. LANCE¹

RÉSUMÉ

Le registre des adresses est une base de sondage d'adresses résidentielles pour les centres urbains de moyenne et de grande dimension qui figurent dans le Fichier principal de région (FPR) de la Division de la géographie de Statistique Canada. Pour la Colombie-Britannique, le registre des adresses a été augmenté afin d'inclure des agglomérations urbaines plus petites ainsi que certaines régions rurales. Dans cet article, on présente un aperçu historique du projet, ses objectifs comme moyen de réduire le sous-dénombrement lors du recensement du Canada de 1991, ses sources et produits, la méthodologie requise pour sa mise en application initiale, l'évaluation postcensitaire proposée et des perspectives pour l'avenir.

MOTS CLÉS: Registre des adresses; sous-dénombrement du recensement; systèmes d'information géographique (SIG).

1. INTRODUCTION

Le concept d'un registre des adresses à Statistique Canada remonte aux années 60. Fellegi et Krötki (1967) ont, pour la première fois, considéré la création d'un de ces registres pour le recensement de 1971, en se basant sur des fichiers administratifs. Leur méthode était surtout manuelle et a donné un ensemble très complet d'adresses avec un sous-dénombrement et un surdénombrement minimum. Au milieu des années 70 (Booth 1976), l'idée est réapparue lors de la planification du recensement de 1981. Cette fois, la méthode utilisée commençait avec la saisie des adresses recueillies au cours du recensement précédent et on ajoutait à ces données des renseignements fournis par Postes Canada. Dans les deux cas, les listes d'adresses produites étaient considérées comme une base de sondage pour un recensement avec envoi par la poste. Cependant, les coûts de création étaient élevés et, pour être efficaces, ils auraient entraîné des réductions compensatrices dans d'autres opérations du recensement. De plus, on considérait que les risques associés au fait de changer la méthode traditionnelle de dénombrement étaient trop élevés. C'est pourquoi, la production d'un registre des adresses a été suspendue dans chaque cas.

Il y a eu un renouveau d'intérêt pour le concept de registre des adresses à la suite de la Conférence internationale sur la planification du recensement de 1991 (Royce 1986, 1987) qui s'est tenue en octobre 1985. Cet intérêt découlait de la possibilité d'automatiser la méthode de Fellegi et Krötki suite aux progrès technologiques, comme la disponibilité de fichiers administratifs, renfermant les adresses et les codes postaux, sur support exploitable par une machine et l'élaboration de logiciels internes pour décomposer les adresses en éléments standard, pour affecter les codes postaux et pour coupler ces derniers aux éléments géographiques du recensement. Cet intérêt découlait aussi de l'élaboration d'une théorie statistique portant sur le couplage d'enregistrements (Fellegi et Sunter 1969) et de systèmes informatiques basés sur cette théorie (Hill et Pring-Mill 1985).

Suite à cet intérêt, un projet a été lancé en 1986, la première recherche (Gamache-O'Leary et coll. 1987) portant sur l'étude de l'utilisation d'un registre des adresses pour un recensement

¹ L. Swain et B. Lafrance, Division des méthodes d'enquêtes sociales; J.D. Drew, Ontario, Canada, K1A 0T6.
K. Lance, Division de la géographie, Statistique Canada, Ottawa, Canada, K1A 0T6.

JUDGE, G.G., et BOCK, M.E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam-New York-Oxford: North-Holland Publishing Company.

ROYCE, D., et LUC, M. (1990). Recalculation of Fellegi's test statistics on census adjustment for the 1981 and 1986 censuses. Rapport interne, Statistique Canada.

ROYCE, D. (1991). Technical criteria for adjusting the population estimates program for census coverage error. Rapport interne, Statistique Canada.

ROYCE, D. (1992). Incorporating estimates of census coverage error into the Canadian population estimates program. *Proceedings of the Eighth Annual Research Conference*, Bureau of the Census, Washington, DC (à paraître).

SAWA, T., et HIROMATSU, T. (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica*, 41, 1093-1101.

SHIRM, A.L., et PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 400, 965-978.

SPENCER, B. (1980). Implications of equity and accuracy for undercount adjustment: A decision-theoretic approach. Dans *Proceedings of the 1980 Conference on Census Undercount*, Bureau of the Census, Washington, DC.

SPENCER, B. (1986). Conceptual issues in measuring improvement in population estimates. Dans *Proceedings of the Second Annual Research Conference*, Bureau of the Census, Washington, DC, 393-407.

TORO-VIZCORRONDO, C., et WALLACE, T.D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 322, 558-572.

WOLTER, K.M., et CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 414, 278-284.

En deuxième lieu, il faut aussi pousser plus loin l'analyse de la sensibilité des résultats par rapport à différents poids dans la fonction de perte. Les résultats de la section 3 avaient été obtenus à la suite de l'utilisation d'un poids égal à l'inverse de l'effectif recensé ou de l'effectif recensé corrigé pour chaque province. Si la pondération avait été différente, les résultats ne seraient pas les mêmes. Un poids plus général qui pourrait être intéressant est $w_k = Y_k^r$, où Y est une sorte de paramètre de puissance. On pourrait alors étudier la sensibilité des résultats de la section 3 par rapport à ce paramètre.

Finalement, bien que les méthodes que nous venons de décrire dans cet article constituent un ensemble organisé pour l'élaboration et l'évaluation d'estimateurs, la manière exacte dont elles seront utilisées reste à définir. Voici les questions particulières qu'il faut résoudre :

1. Quelle est l'importance relative de diverses fonctions telles que les totaux, les proportions et les taux de croissance? Des fonctions différentes donneront des résultats différents mais en définitive, il faut choisir un seul estimateur afin d'assurer la cohérence des estimations.
2. À quels niveaux d'aggrégation géographique et démographique ces méthodes devraient-elles être utilisées? Par exemple, devrait-on utiliser l'estimateur de test préliminaire ou l'estimateur composite décrits dans la section 3 au niveau provincial, au niveau provincial croisé avec les groupes d'âge-sexe ou à des niveaux encore plus fins? Les résultats obtenus dépendent du niveau d'analyse utilisé.
3. Pourrions-nous même envisager d'utiliser les estimateurs composites comme estimateurs de premier plan, par exemple comme estimateurs de la population totale des provinces? Il serait peut-être difficile d'expliquer aux utilisateurs pourquoi les redressements ne concordent pas avec les estimations officielles du sous-dénombrement.

Comme il faut faire preuve de jugement professionnel pour résoudre de telles questions, la décision d'opérer ou non un redressement ne peut être prise de façon automatique suivant des critères établis (il en va de même du choix de la méthode de redressement). Tandis que les méthodes que nous avons décrites dans cet article peuvent à coup sûr servir de jalons, la décision finale devra reposer sur un examen sérieux où l'on mettra en balance les possibilités de gain dans la précision des estimations et la facilité avec laquelle les méthodes choisies pourront être communiquées aux utilisateurs des données du PED et comprises par ceux-ci.

REMERCIEMENTS

L'auteur tient à remercier le rédacteur en chef, les deux arbitres ainsi que Richard Carter pour les nombreux commentaires utiles qui ont contribué à rehausser la qualité de cet article.

BIBLIOGRAPHIE

ANDREWS, D. (1991). Discussion sollicitée lors de la réunion du comité consultatif des méthodes statistiques de Statistique Canada, octobre 1991.

BROOK, R.J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association*, 71, 353, 126-131.

CITRO, C.F., et COHEN, M.L. (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.

COHEN, A. (1965). Estimates of the linear combination of parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics*, 36, 78-87.

FELLECI, I.P. (1980). Should the census count be adjusted for allocation purposes? Equity considerations. Dans *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census.

où

$$EPCMP(\bar{g}(\bar{T}_1, \bar{g}(\bar{T}_2))) = \sum_{ij}^U w_{ij} [Cov(U_{1i}, U_{2j}) + Biases(U_{1i}) Biases(U_{2j})] \tag{27}$$

est défini comme l'erreur de produit croisé moyenne pondérée de $\bar{g}(\bar{T}_1)$ et $\bar{g}(\bar{T}_2)$. L'EPCMP de notre estimateur composite est minimisée lorsque

$$\alpha = \frac{EPCMP(\bar{g}(\bar{T}_2)) - EPCMP(\bar{g}(\bar{T}_1), \bar{g}(\bar{T}_2))}{EPCMP(\bar{g}(\bar{T}_1)) + EPCMP(\bar{g}(\bar{T}_2)) - 2EPCMP(\bar{g}(\bar{T}_1), \bar{g}(\bar{T}_2))} \tag{28}$$

Pour obtenir une estimation de α , on remplace l'EPCMP et l'EPCMP dans l'équation ci-dessus par leurs valeurs estimées.

Pour illustrer l'utilisation de cette méthode, supposons qu'on a décidé de redresser l'effectif de la population d'une province. Pour effectuer le redressement à tous les niveaux d'agrégation, on a le choix entre le redressement général (c.-à-d. redresser toutes les quantités infra-provinciales au moyen du même facteur) et le redressement synthétique, où les corrections se font séparément dans plusieurs groupes d'âge-sexe. L'avantage de la première méthode est qu'elle utilise uniquement l'estimation du sous-dénombrement pour l'ensemble de la province, laquelle estimation est vraisemblablement plus fiable que les estimations du sous-dénombrement par groupe d'âge-sexe au niveau provincial. En revanche, si le sous-dénombrement varie beaucoup entre les groupes d'âge-sexe et si les quantités infra-provinciales désignées par l'indice i varient elles aussi selon la composition âge-sexe, il sera préférable d'opter pour l'estimateur synthétique.

S'il est possible d'obtenir les valeurs estimées de U_i d'une source quelconque, on peut estimer toutes les composantes (covariance et biais) des EPCMP et de l'EPCMP (en se servant de formules comme celles du tableau 3) et déterminer l'estimateur composite optimal formé de l'estimateur général et de l'estimateur synthétique. Bien qu'en pratique, il soit presque impossible de connaître les U_i pour des quantités infra-provinciales, on peut envisager l'application de la méthode à des niveaux supérieurs. Par exemple, on pourrait assimiler les provinces aux quantités désignées par l'indice i et se servir de facteurs de redressement (général et synthétique) calculés pour l'ensemble du Canada. Une deuxième solution serait de créer une population artificielle (comme dans Shirm et Preston (1987) ou Wolter et Causey (1991) par exemple) où les U_i sont supposés connus.

5. SUJETS DE RECHERCHE

L'étude que nous venons d'exposer n'est qu'une première étape dans l'analyse et la comparaison du rendement de divers estimateurs d'une série de chiffres de population. Beaucoup de recherches restent encore à faire sur de nombreux sujets.

Premièrement, il faut étudier plus en profondeur l'EPCMP de l'estimateur de test préliminaire et de l'estimateur composite pour les cas plus généraux décrits dans les sections 3 et 4. Bien que l'on ne soit pas encore parvenu à définir des expressions analytiques pour ces EPCMP, il se pourrait que les résultats plus généraux relatifs aux estimateurs de test préliminaire et aux estimateurs de Stein-Rule dont font état Judge et Bock (1978) s'avèrent utiles. Si tel était le cas, nous pourrions nous servir de ces résultats pour répondre à des questions comme celles-ci: Peut-on trouver des valeurs critiques optimales pour les estimateurs de test préliminaire des sections 3.3 et 4.1 suggérées par Fellegi? Comment l'EPCMP de l'estimateur de test préliminaire se compare-t-elle dans la pratique à celle des trois autres estimateurs?

4.1 Estimateurs de test préliminaire

Tout comme ce fut le cas dans les sections 2 et 3, nous pouvons nous servir de l'EOMP pour construire des tests statistiques qui permettront de choisir l'un ou l'autre de deux estimateurs. Supposons, par exemple, que nous devions choisir entre l'estimateur fondé sur les chiffres non redressés du recensement et l'estimateur général pour des totaux de population (le redressement "général" n'a évidemment aucun effet sur les proportions de population). Si nous comparons les EOMP de ces deux estimateurs, nous constatons que l'estimateur général sera préférable à l'estimateur fondé sur les chiffres du recensement si

(24)
$$\sigma^2 < U^2(1 - R^2) \left[1 - \frac{2TB}{U(1 - R)} \right],$$

où

(25)
$$B = 1 - \frac{\sum_{i=1}^I \frac{\lambda_i^2}{\tau_i}}{1}$$

et $\tau_i = T_i/T$. Cette condition est définie sous une forme différente dans Wolter et Causey (1991). B est un indice de l'hétérogénéité du sous-dénombrement; il est non négatif, et est égal à zéro si et seulement si le sous-dénombrement est parfaitement uniforme. L'inéquation (24) étant semblable à l'inéquation (1) (si l'on fait abstraction du terme entre crochets), il est possible de construire un test très comparable à celui qui est décrit dans la section 2.3. La valeur critique du coefficient de variation dépendra, comme avant, du niveau de signification choisi et du biais relatif mais aussi de B/U , le rapport entre l'indice d'hétérogénéité du sous-dénombrement et le taux de sous-dénombrement global. Royce (1991) a montré que ce facteur additionnel a un effet plutôt négligeable en pratique sur le CV critique. Par conséquent, si le redressement est justifié aux niveaux supérieurs, nous ne voyons pas pourquoi il ne le serait pas à des niveaux inférieurs. Wolter et Causey (1991) arrivent à des conclusions semblables dans une étude de simulation.

4.2 Estimateurs composites

Dans les sections 2 et 3, nous avons examiné des estimateurs composites dont les deux versions extrêmes correspondaient l'une, à l'estimateur fondé sur les chiffres non redressés du recensement et l'autre, à l'estimateur fondé sur les chiffres redressés des estimateurs composites de l'estimateur synthétique et de l'estimateur général, l'éventail des estimateurs composites s'est accru sensiblement. Par exemple, on pourrait envisager des estimateurs composites d'un estimateur fondé sur les chiffres non redressés et d'un estimateur synthétique, d'un estimateur fondé sur les chiffres redressés et d'un estimateur général, d'un estimateur synthétique, et ainsi de suite. Par conséquent, nous décrivons ci-dessous une méthode qui permet de construire un estimateur composite formé de deux estimateurs quelconques.

L'estimateur composite général est défini par l'expression $\overline{U}^* = \alpha \overline{U}_1 + (1 - \alpha) \overline{U}_2$ où \overline{U}_1 et \overline{U}_2 sont deux estimateurs. L'EOMP de cet estimateur est

$$EOMP(\overline{g}(\overline{T}^*)) = \alpha^2 EOMP(\overline{g}(\overline{T}_1)) + (1 - \alpha)^2 EOMP(\overline{g}(\overline{T}_2))$$

(26)
$$+ 2\alpha(1 - \alpha) EPCMP(\overline{g}(\overline{T}_1, \overline{g}(\overline{T}_2))),$$

4. ESTIMATION POUR PETITES RÉGIONS

Dans les deux sections précédentes, nous avons vu qu'il était possible d'obtenir des estimations directes du sous-dénombrement, ainsi que les variances estimées correspondantes, grâce aux études de mesure de la couverture. Cela est possible notamment pour les provinces, pour certaines grandes régions métropolitaines de recensement et pour de grands groupes démographiques (ex.: âge-sexe, âge-état matrimonial) au niveau national. Or, le Programme des estimations démographiques (PED) produit des estimations à des niveaux de détail très poussés (ex.: par année d'âge, sexe et état matrimonial pour quelque 260 divisions de recensement). En règle générale, il n'existe pas d'estimations directes du sous-dénombrement à ces niveaux. Néanmoins, à cause de la nécessité d'assurer une cohérence dans les estimations, il faut veiller à ce que tout redressement effectué aux niveaux supérieurs "se répercute" jusqu'aux niveaux les plus bas utilisés dans le PED. Dans cette section, nous voyons comment l'estimation synthétique peut être utile dans cette perspective et comment on peut encore se servir de l'EQMP pour construire des estimateurs de test préliminaire et des estimateurs composites.

L'estimation synthétique repose sur l'hypothèse que le sous-dénombrement net est uniforme à l'intérieur d'un certain nombre de "groupes de redressement" désignés par l'indice a . L'estimation synthétique est calculée au moyen de la formule $U_i^s = \sum^a \lambda_{ia} U_a^a$ où $\lambda_{ia} = Y_{ia}/Y^a$. Par exemple, les groupes de redressement pourraient être les groupes d'âge-sexe, pour lesquels il existe des estimations du sous-dénombrement U_a^a à un niveau supérieur.

Lorsqu'il y a un seul groupe de redressement, nous sommes devant un cas particulier de l'estimation synthétique que Wolter et Causey (1991) appellent l'estimateur général. Il est défini par la formule $U_i^{ATB} = \lambda_i U$ où $\lambda_i = Y_i/Y$. On peut calculer l'EQMP de l'estimateur général et de l'estimateur synthétique au moyen de l'équation (14). Comme w_{ij} ne dépend pas de l'estimateur utilisé, seule l'expression entre crochets varie. Dans le tableau 3, nous comparons les estimateurs de U_i ainsi que les termes de covariance et de biais correspondants pour les quatre estimateurs suivants: estimateur fondé sur les chiffres bruts du recensement, estimateur fondé sur les chiffres redressés, estimateur général et estimateur synthétique.

Tableau 3
Exemples de covariances et de biais dans la formule d'approximation de l'EQMP pour divers estimateurs

Estimateur	U_i^*	$\text{Cov}(U_i^*, U_j^*)$	Biais (U_i^*)
Chiffres bruts du recensement	0	0	$-U_i$
Chiffres redressés du recensement	U_i	σ_{ij}	b_i
Général	$\lambda_i U$	$\lambda_i \lambda_j \sigma^2$	$\lambda_i(U + b) - U_i$
Synthétique	$\sum^a \lambda_{ia} U_a^a$	$\sum^{aa'} \lambda_{ia} \lambda_{ja'} \sigma_{aa'}$	$\sum^a \lambda_{ia}(U_a^a + b_a) - U_i$

où $b = \sum^i b_i$ et de même, b_a est le biais de U_a^a .

Dans le cas des totaux de population par exemple, le degré de redressement estimé est

$$\alpha_L = \frac{\sum_i T_i \hat{U}_i^2}{\sum_i T_i [\hat{\sigma}_i^2 + \hat{U}_i^2]}, \tag{22}$$

où \hat{U}_i est le taux de sous-dénombrement estimé, c.-à-d. $\hat{U}_i/(Y_i + U_i)$, et $\hat{\sigma}_i^2$, la variance estimée correspondante.

En ce qui concerne les proportions de population, le degré de redressement est calculé au moyen de la formule

$$\alpha_L = \frac{\sum_i T_i [\hat{\sigma}_i^2 + \hat{U}_i^2] - T(\hat{\sigma}^2 + \hat{U}^2)}{\sum_i T_i \hat{U}_i^2 - T\hat{U}^2}, \tag{23}$$

où \hat{U} est le taux de sous-dénombrement estimé pour la population globale, c.-à-d. $\sum_i U_i/\sum_i (Y_i + U_i)$, et $\hat{\sigma}^2$, la variance estimée correspondante. L'inverse de l'effectif recensé corrigé a servi de poids dans ces deux exemples.

3.5 Comparaisons numériques

Pour les totaux de population uniques, il a été possible de définir des formules exactes ou approximatives pour l'EQM des quatre estimateurs étudiés, en fonction de U^2/σ^2 , de R , de r et (dans le cas de l'estimateur de test préliminaire) de la valeur critique du test. Malheureusement, il n'est pas encore possible de faire de même pour l'EQMP de fonctions complexes d'un vecteur de totaux de population.

Néanmoins, on peut estimer l'EQMP de trois estimateurs – celui fondé sur les chiffres non redressés du recensement, celui fondé sur les chiffres redressés et l'estimateur composite – en substituant dans l'équation (18) des estimations du sous-dénombrement ainsi que les variances estimées correspondantes (s'il existe des estimations du biais, on peut aussi s'en servir; cependant, nous supposons ici qu'il n'y a pas de biais). À titre d'exemple, les figures 8 et 9 contiennent un graphique où le rapport estimé de l'EQMP à l'EQMP optimale pour le recensement de 1981 est exprimé en fonction d' α ; la encore, les provinces sont les unités désignées par i . Les valeurs extrêmes de $\alpha = 0$ et $\alpha = 1$ correspondent respectivement au cas des chiffres non redressés du recensement et au cas des chiffres redressés, tandis que le point minimum de la courbe correspond à l'alpha optimal. La figure 8 porte sur les totaux et la figure 9 sur les proportions. Les valeurs optimales d' α ont été calculées au moyen des formules (22) et (23). Dans chaque cas, le degré de redressement optimal est proche de 1.0 et donne une EQMP beaucoup moins élevée que celle qui correspond à l'absence de redressement (par un facteur de près de 70, par exemple, en ce qui concerne les totaux de population). Le degré de redressement optimal est moins élevé pour les proportions que pour les totaux, ce qui dénote une fois de plus que les estimations de différences de taux de sous-dénombrement entre les provinces sont moins précises que les estimations de taux de sous-dénombrement. Il est intéressant aussi de constater que l'EQMP qui correspond au redressement complet n'est que légèrement plus élevée que celle qui correspond au degré de redressement optimal. Cette similitude peut avoir des conséquences pratiques importantes car il est beaucoup plus facile de justifier un redressement complet qu'un redressement partiel devant les utilisateurs de données.

Tableau 2

Valeurs z des tests de Fellegi pour le redressement des totaux et des proportions de population par province, Contre-vérification des dossiers, 1976, 1981 et 1986

Fonction	1976	1981	1986
Totaux	9.3	10.1	13.1
Proportions	3.1	1.8	1.5

3.4 Estimateur composite

À première vue, nous serions portés à croire que $\alpha_i U_i$ est une extension naturelle de l'estimateur composite de la section 2.4. Or, la variation du degré de redressement selon la valeur de i cause des problèmes de cohérence. Par exemple, nous pourrions devoir opérer un redressement plus profond au niveau national qu'au niveau provincial du fait que les estimations du sous-dénombrement pour les provinces sont moins précises que celles pour le Canada. Si nous devons faire cela, la somme des totaux provinciaux ne concorderait pas avec le total pour le Canada.

Par conséquent, nous nous limitons dans la pratique à une seule valeur de alpha, c.-à-d. $\bar{Q}^\alpha = \alpha \bar{Q}$, où, là encore, $0 \leq \alpha \leq 1$. L'EOMP de cet estimateur est

$$EOMP(\bar{Q}^\alpha) = \sum_{ij} w_{ij} [\alpha^2 (\sigma_{ij}^2 + b_i b_j) + (\alpha - 1)^2 U_i U_j + 2\alpha (\alpha - 1) U_i b_j],$$

(18)

laquelle expression est minimisée lorsque

$$\alpha = \frac{\sum_{ij} w_{ij} U_i (U_j + b_j)}{\sum_{ij} w_{ij} [\sigma_{ij}^2 (U_i + (U_i + b_i)(U_j + b_j))]}.$$

(19)

Si, comme dans la section 3.3, nous posons l'hypothèse que $\sum_{ij} w_{ij} b_i (U_j + b_j) \leq 0$ une borne inférieure pour l'alpha optimal est définie

$$\alpha = \frac{\sum_{ij} w_{ij} (U_i + b_i)(U_j + b_j)}{\sum_{ij} w_{ij} [\sigma_{ij}^2 (U_i + (U_i + b_i)(U_j + b_j))]}.$$

(20)

que nous estimons par

$$\alpha = \frac{\sum_{ij} w_{ij} \sigma_{ij}^2}{\sum_{ij} w_{ij} [\sigma_{ij}^2 (\sigma_{ij}^2 + b_i b_j)]}.$$

(21)

en supposant que les w_{ij} sont connus. En pratique, on estime les w_{ij} en substituant dans l'équation (15), comme on l'a fait pour l'estimateur de test préliminaire, les chiffres bruts du recensement ou les chiffres redressés.

3.1 Estimateur fondé sur les chiffres non redressés du recensement

L'EOMP de l'estimateur fondé sur les chiffres non redressés du recensement est $EOMP(\hat{Q}^c) = \sum_{ij} \omega_{ij} U_i U_j$.

3.2 Estimateur fondé sur les chiffres redressés du recensement

L'EOMP de l'estimateur fondé sur les chiffres redressés est $EOMP(\hat{Q}^A) = \sum_{ij} \omega_{ij} [\sigma_{ij} + b_i b_j]$ où $\sigma_{ij} = Cov(U_i, U_j)$ et $b_i = Biais(U_i)$.

3.3 Estimateur de test préliminaire

Comme dans la section 2.3, nous préférons l'estimateur fondé sur les chiffres redressés à l'estimateur fondé sur les chiffres non redressés si l'EOMP du premier est moindre que celle du second, c.-à-d. si

(16)
$$D = \sum_{ij} \omega_{ij} [U_i U_j - \sigma_{ij} - b_i b_j] > 0.$$

Fellegi (1980) a proposé des tests pour ce genre d'hypothèses dans le cas particulier des totaux de population et des proportions de population; cependant, ces notions peuvent s'appliquer facilement à n'importe quelle fonction \bar{g} . Le membre de gauche de l'inéquation (16) est estimé par $\hat{D} = \sum_{ij} \omega_{ij} [U_i U_j - 2\sigma_{ij}]$ où les ω_{ij} sont supposés connus. En pratique, on estime les ω_{ij} en substituant dans l'équation (13) les chiffres bruts du recensement ou les chiffres redressés. Fellegi soutient qu'une faible variation des poids sera peu susceptible de modifier sensiblement les résultats des tests. On peut alors montrer facilement que $E(\hat{D}) = D + 2 \sum_{ij} \omega_{ij} b_i (U_j + b_j)$. En ce qui concerne les totaux et les proportions, Fellegi a présenté des arguments qui expliquent pourquoi on peut supposer que le second terme du membre de droite est non positif, c.-à-d. $\sum_{ij} \omega_{ij} b_i (U_j + b_j) \leq 0$ de sorte que \hat{D} tendrait à sous-estimer D . Fellegi a aussi calculé une variance approximative pour \hat{D} . Grâce à cela et à l'hypothèse que \hat{D} est distribué normalement, on a pu construire un test pour l'hypothèse posée en (16).

Dans le cas plus général, la variance approximative de \hat{D} est définie $Var(\hat{D}) = 4 \sum_{ij} \omega_{ij} \sigma_{ij} (\sum_{i'j'} \omega_{i'j'} \omega_{ij} U_{i'} U_{j'})$. On peut alors calculer une estimation de $Var(\hat{D})$ en remplaçant U_i et σ_{ij} dans l'équation ci-dessus par leurs valeurs estimées.

Dans le cas des totaux par exemple, la variable à tester (valeur z) est définie

(17)
$$z = \frac{\hat{D}}{\sqrt{Var(\hat{D})}} = \frac{\sum_i \frac{Q_i^2 - 2\sigma_i^2}{X_i}}{\sqrt{\sum_i \frac{Q_i^2 \hat{\sigma}_i^2}{Y_i^2}}},$$

où, en l'occurrence, l'inverse de l'effectif recensé sert de poids. Une expression semblable existe pour les proportions de population.

Le tableau 2 donne les valeurs z calculées pour les totaux et les proportions de population par province pour les recensements de 1976, 1981 et 1986. Les données sont tirées des contre-vérifications de dossiers effectuées lors de ces recensements.

D'après ce tableau, le redressement des totaux de population prime largement le redressement des proportions de population, ce qui dénote que les estimations de différences de taux de sous-dénombrement entre les provinces sont moins précises que les estimations de taux de sous-dénombrement. D'autres résultats numériques figurent dans Royce et Luc (1990).

Cette formule a ceci d'utile qu'elle exprime chaque composante de la fonction de risque en deux parties: un poids, ω_{ij} , qui dépend uniquement de w_k et de la fonction \bar{g} , et l'expression entre crochets, qui dépend uniquement de l'estimateur utilisé.

Tandis que le choix de w_k peut être arbitraire, des considérations d'équité ont souvent amené les expérimentateurs à choisir $w_k = 1/T_k$. En ce qui concerne les totaux et les proportions de population par exemple, la fonction de risque devient alors l'équivalent des fonctions proposées par Fellegi (1980) et utilisées par Wolter et Causey (1991), pour ne nommer que ceux-ci. Les autres valeurs proposées pour w_k dans les ouvrages statistiques sont $w_k = 1/Y_k$, $w_k = 1/\bar{T}_k$, et $w_k = 1$. Pour une analyse plus détaillée des avantages de ces divers poids, le lecteur est prié de consulter les ouvrages mentionnés plus haut. Le tableau 1 donne des exemples de poids ω_{ij} pour diverses fonctions.

En ce qui concerne les taux de croissance démographique, la première paire d'indices désigne la population étudiée (par ex.: population d'une province), tandis que la seconde paire d'indices désigne le recensement de la période 1 et de la période 2 respectivement. Le second indice rattaché à T_i désigne aussi le recensement de la période 1 ou de la période 2.

Dans le reste de cette section, nous montrons comment l'EOMP sert à l'élaboration et à l'évaluation de nos quatre estimateurs: estimateur fondé sur les chiffres non redressés du recensement, estimateur fondé sur les chiffres redressés, estimateur de test préliminaire et estimateur composite.

Tableau 1

Exemples de poids ω_{ij} dans la formule d'approximation de l'EOMP pour diverses fonctions

Fonction	Série de totaux de population	Série de proportions de population	Série de taux de croissance
ω_{ij}	$w_{ii} = w_i$ $\omega_{ij} = 0 \quad i \neq j$	$\omega_{ii} = \frac{1}{T_i} \left(\sum_k w_k T_k^2 + w_i T_i^2 - 2w_i T T_i \right)$ $\omega_{ij} = \frac{1}{T_i} \left(\sum_k w_k T_k^2 - T(w_i T_i + w_j T_j) \right) \quad i \neq j$	$\omega_{iii1} = \frac{T_i^4}{w_i T_i^2}$ $\omega_{iii2} = - \frac{T_i^4}{w_i T_i T_i^2} = \omega_{iii1}$ $\omega_{iij2} = \frac{T_i^4}{w_i T_i^2} = \omega_{iij1}$ $\omega_{ij11} = \omega_{ij12} = \omega_{ij21} = \omega_{ij22} = 0 \quad i \neq j$

Les figures 5, 6 et 7 contiennent un graphique où l'EQM de l'estimateur composite est exprimée en fonction de U^2/σ^2 ; dans ces mêmes graphiques sont reproduites aussi les EQM de l'estimateur fondé sur les chiffres non redressés du recensement, de l'estimateur fondé sur les chiffres redressés et de l'estimateur de test préliminaire optimal de la section 2.3. Lorsqu'il n'y a pas de biais (figure 5) ou que le biais est positif (figure 6), l'estimateur composite est supérieur à l'estimateur de test préliminaire optimal. Par contre, lorsque le biais est négatif (figure 7), l'EQM de l'estimateur composite peut être beaucoup plus élevée que celle des autres estimateurs pour une portion considérable de l'intervalle de valeurs de U^2/σ^2 .

3. ESTIMATEURS À CARACTÈRE PLUS GÉNÉRAL

Dans cette section, nous généralisons de deux manières les quatre estimateurs étudiés dans la section précédente. Premièrement, au lieu de ne considérer que des totaux de population uniques, nous allons étudier des vecteurs de totaux de population, désignés par $\bar{T} = (T_1, T_2, \dots, T_N)$. Deuxièmement, nous ne nous limitons pas aux totaux de population; nous considérons aussi les fonctions de ces totaux, désignées par $\bar{g}(\bar{T}) = (g_1(\bar{T}), g_2(\bar{T}), \dots, g_K(\bar{T}))$, où, de façon générale, $K \neq N$. Les principales fonctions de totaux de population comprennent les proportions de population, utilisées pour les transferts de fonds entre l'administration fédérale et les administrations provinciales, les taux de croissance entre les recensements, les différences de taux de croissance entre les provinces, et ainsi de suite.

En évaluant la précision globale d'une estimation $\bar{g}(\bar{T}^*)$ de $\bar{g}(\bar{T})$, nous allons recourir à une fonction de perte. L'utilisation des fonctions de perte dans le but d'évaluer les effets du redressement des chiffres du recensement est traitée dans Fellegi (1980), Citro et Cohen (1985), Spencer (1986) et Wolter et Causey (1991), pour n'en nommer que quelques-uns. La fonction de perte qui est utilisée ici est une généralisation des fonctions de perte proposées antérieurement pour des totaux et des proportions de population. De façon plus précise, le risque (perte moyenne espérée) de l'estimateur $\bar{g}(\bar{T}^*)$ est l'erreur quadratique moyenne pondérée (EQMP), qui est définie par l'expression

$$(13) \quad EQMP(\bar{g}(\bar{T}^*)) = E \left\{ \sum_{k=1}^K w_k (g_k(\bar{T}^*) - g_k(\bar{T}))^2 \right\},$$

où w_k est un poids défini par l'utilisateuse qui indique l'importance de la k -ième composante de la fonction de perte. Comme \bar{g} peut être complexe dans la pratique, il est utile de disposer aussi d'une approximation de l'EQM obtenue en développant $\bar{g}(\bar{T}^*)$ en une série de Taylor par rapport à \bar{T} . On a ainsi:

$$(14) \quad EQMP \bar{g}(\bar{T}^*) \doteq \sum_N \sum_{j=1}^I w_{ij} [Cov(U_i^*, U_j^*) + Biais(U_i^*) Biais(U_j^*)]$$

où le poids w_{ij} est défini par l'expression

$$(15) \quad w_{ij} = \sum_{k=1}^K w_k \frac{\partial T_i}{\partial g_k} \frac{\partial T_j}{\partial g_k}.$$

(Notons que l'approximation de l'EQM peut aussi être formulée comme l'espérance mathématique de la forme quadratique $(\bar{Q}^* - \bar{U})' \Omega(\bar{Q}^* - \bar{U})$, où w_{ij} est l'élément ij de Ω .)

Dans Judge et Bock (1978), on trouve d'autres méthodes pour choisir la valeur optimale de c , par exemple la minimisation de la distance moyenne (au lieu de l'écart maximum) et des méthodes bayésiennes.

2.4 Estimateur composite

Nous avons vu que l'estimateur de test préliminaire s'écrit $U^P = IU$, où I ne peut prendre que les valeurs 0 ou 1. Or, on a montré que cet estimateur est inacceptable (Cohen, 1965) à cause justement de cette discontinuité inhérente. Dans ces circonstances, le multiplicateur de U pourrait peut-être prendre n'importe quelle valeur entre 0 et 1. Autrement dit, au lieu de nous servir des données pour savoir s'il faut procéder ou non à un redressement, nous en servons pour savoir quelle doit être la mesure du redressement. Spencer (1980) et, plus récemment, Andrews (1991) ont déjà parlé de ce type d'estimateur. Posons $U^\alpha = \alpha U$ où $0 \leq \alpha \leq 1$. Pour une valeur α donnée, l'estimateur ci-dessus a une EQM égale à

$$\text{EQM}(\alpha U) = \alpha^2 \sigma^2 + U^2 (\alpha(1 + R) - 1)^2, \tag{6}$$

laquelle expression est minimisée lorsque

$$\alpha = \frac{U^2(1 + R)^2}{U^2(1 + R)^2 + U^2(1 + R)^2}. \tag{7}$$

Si σ est supposé connu, un estimateur possible de α est

$$\hat{\alpha} = \frac{U^2}{U^2 + U^2(1 + R)^2} \tag{8}$$

et par conséquent,

$$U^{\hat{\alpha}} = \frac{U^3}{U^3 + U^2(1 + R)^2}. \tag{9}$$

On peut calculer l'EQM approximative de cet estimateur au moyen d'une approximation par série de Taylor. En posant

$$h(U, \sigma^2) = \frac{U^3}{(1 + R)(\sigma^2 + U^2)} \tag{10}$$

nous obtenons (en supprimant les termes dont l'ordre est supérieur à celui des termes de dérivée première)

$$\text{EQM}(U^{\hat{\alpha}}) \approx (h(U, \sigma^2) - U)^2 + \left(\frac{\partial h(U, \sigma^2)}{\partial U} \right)^2 (U^2 R^2 + \sigma^2) \tag{11}$$

On peut aussi utiliser cette approximation lorsque σ est inconnu en posant l'hypothèse formulée en (3). La formule de l'EQM comprend alors le terme additionnel:

$$\left(\frac{\partial h(U, \sigma^2)}{\partial \sigma^2} \right)^2 \frac{\sigma^4}{2}. \tag{12}$$

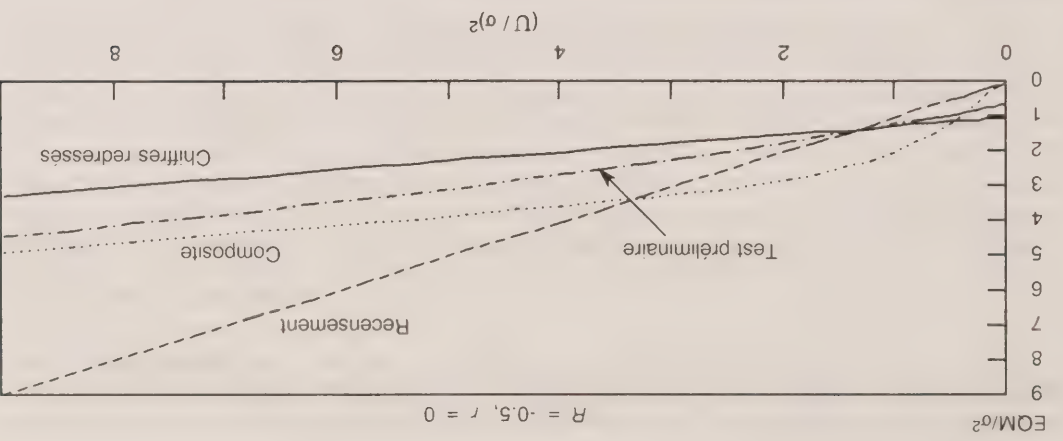


Figure 7 : Comparaison d'EQM

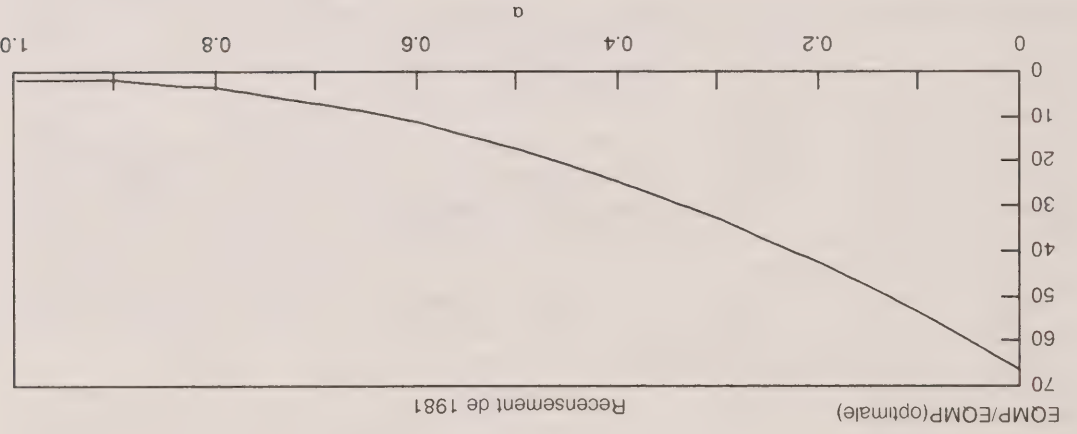


Figure 8 : EQMP pour totaux de population

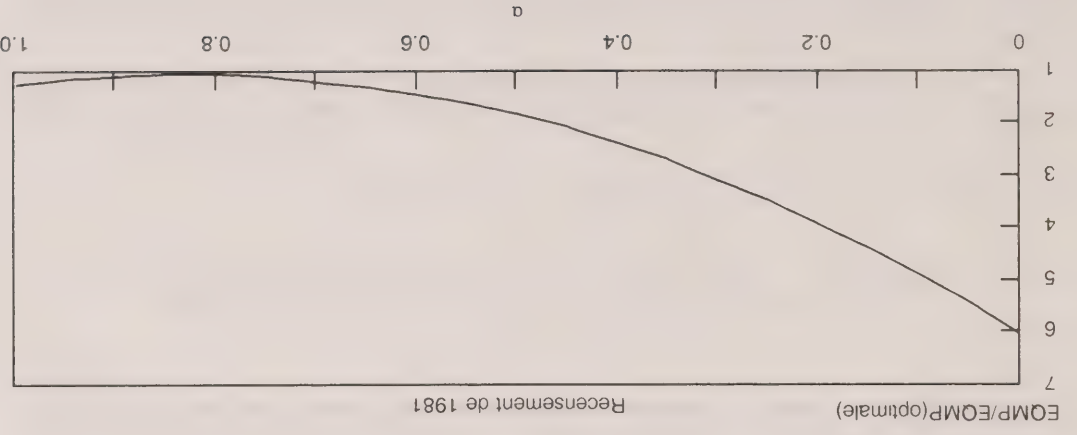
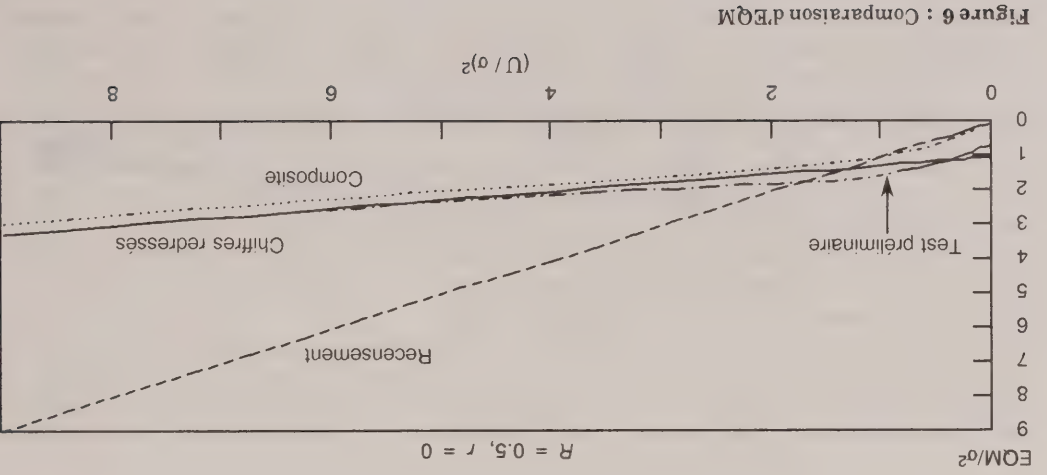
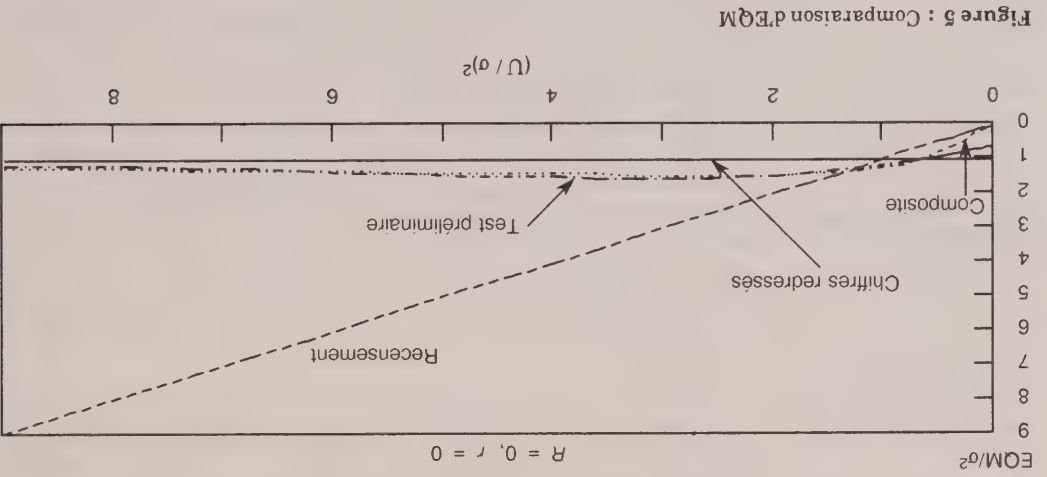
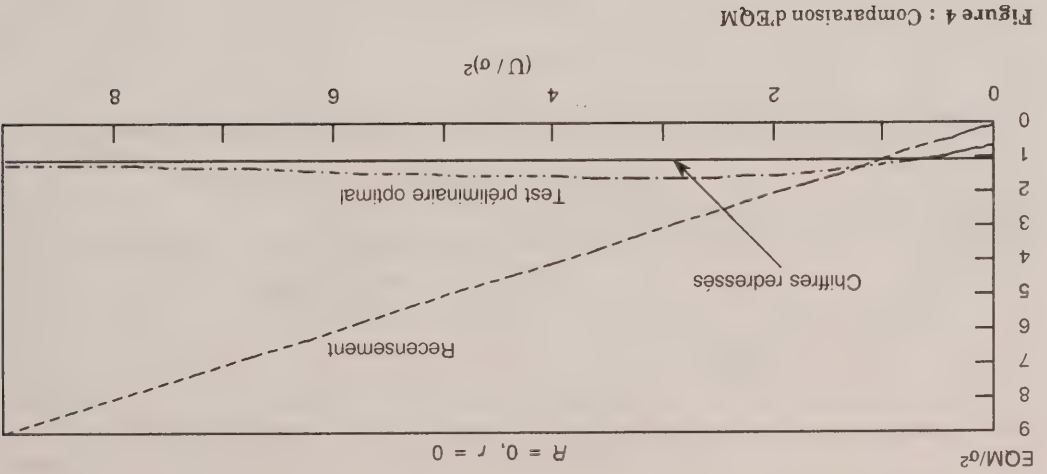


Figure 9 : EQMP pour proportions de population



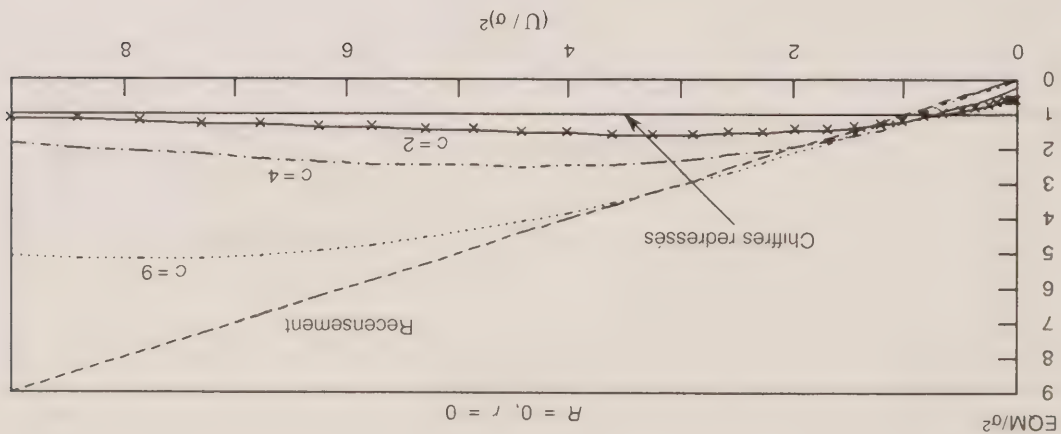


Figure 1 : Comparaison d'EQM

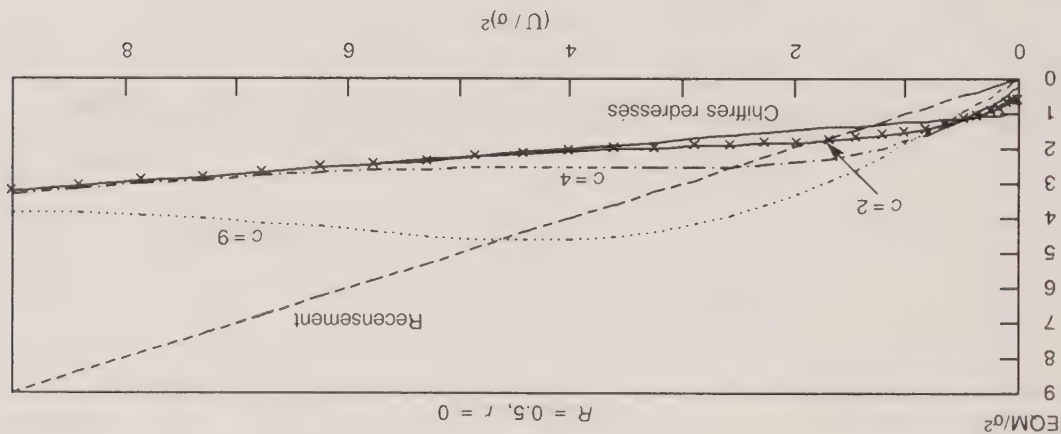


Figure 2 : Comparaison d'EQM

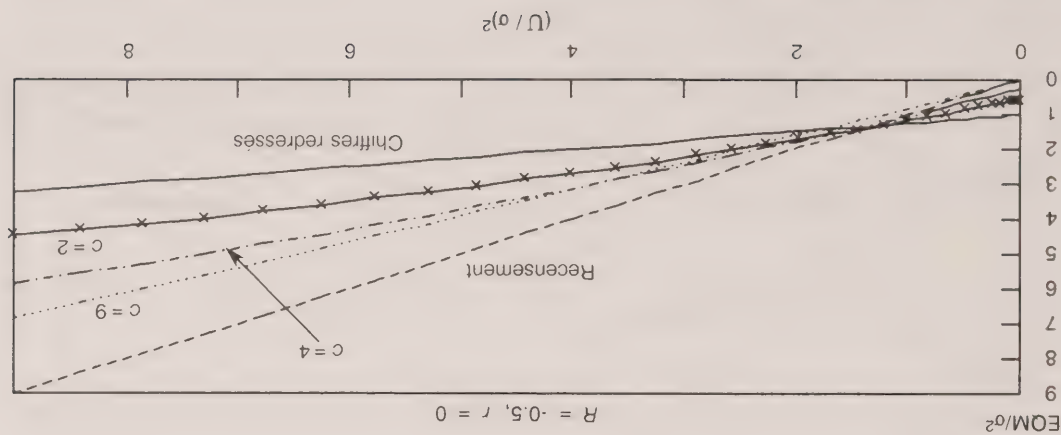


Figure 3 : Comparaison d'EQM

$$I = 1 \quad \text{si} \quad \frac{\sigma^2}{U^2} > c$$

$$= 0 \quad \text{si} \quad \frac{\sigma^2}{U^2} \leq c. \quad (4)$$

Lorsque σ^2 est connu, on peut montrer que l'EQM de cet estimateur est définie (voir, par exemple, Judge et Bock (1978), p. 72)

$$EQM(U^p) = \sigma^2 + U^2 R^2 + (2U^2(1 + R) - \sigma^2) Pr\{\chi^2_{(3,\lambda)} \leq c\} - U^2(1 + R)^2 Pr\{\chi^2_{(5,\lambda)} \leq c\}. \quad (5)$$

Notons que lorsque $c \rightarrow \infty$, c.-à-d. que la probabilité d'un redressement est de plus en plus minime, l'EQM tend vers U^2 , soit l'EQM de l'estimateur fondé sur les chiffres non redressés du recensement. De même, lorsque $c \rightarrow 0$, c.-à-d. que la probabilité d'un redressement est de plus en plus grande, l'EQM tend vers $\sigma^2 + U^2 R^2$, soit l'EQM de l'estimateur fondé sur les chiffres redressés du recensement. Par conséquent, les deux premières méthodes d'estimation présentées dans cette section peuvent être considérées comme des versions extrêmes de la troisième méthode.

La figure 1 renferme un graphique où la valeur EQM/σ^2 pour l'estimateur de test préliminaire est exprimée en fonction de U^2/σ^2 pour diverses valeurs de c dans le cas où il n'y a pas de biais ($R = r = 0$). Les valeurs EQM/σ^2 pour les estimateurs fondés sur les chiffres redressés et les chiffres non redressés du recensement sont aussi reproduites dans ce graphique. Pour chaque valeur de c proposée, la courbe de l'EQM de l'estimateur de test préliminaire débute plus haut sur l'axe des y par rapport à la courbe de l'EQM de l'estimateur fondé sur les chiffres non redressés, croise la courbe de l'EQM de l'estimateur fondé sur les chiffres redressés, atteint un maximum, puis se rapproche de la courbe de l'EQM de l'estimateur fondé sur les chiffres redressés. À mesure que c diminue et que, par voie de conséquence, le niveau de signification α du test augmente, l'EQM de l'estimateur de test préliminaire se rapproche plus rapidement de celle de l'estimateur fondé sur les chiffres redressés; en revanche, elle est plus élevée pour de faibles valeurs de U^2/σ^2 . Par conséquent, le rendement de l'estimateur de test préliminaire pour l'intervalle des valeurs de U^2/σ^2 dépend du niveau de signification choisi pour le test.

Les figures 2 et 3 renferment des graphiques similaires pour les cas où $R = .5$ et $R = -.5$ respectivement (comme nous pourrions croire ne pas disposer de l'information voulue pour établir une valeur estimée de R , nous avons posé $r = 0$). Là encore, l'EQM de l'estimateur de test préliminaire se rapproche de celle de l'estimateur fondé sur les chiffres redressés à mesure que U^2/σ^2 augmente. Elle se rapproche plus rapidement si le biais est positif que s'il n'y a pas de biais mais moins rapidement si le biais est négatif.

Quelle est la 'meilleure' valeur de c pour le test? Idéalement, il faudrait choisir la valeur de c de telle manière que l'EQM de l'estimateur de test préliminaire soit le plus comparable possible à la moins élevée des EQM de l'estimateur fondé sur les chiffres redressés et de l'estimateur fondé sur les chiffres non redressés. À cet égard, Sawa et Hiromatsu (1973) ont proposé une méthode, élargie par la suite par Brook (1976), qui consiste à minimiser l'écart maximum entre l'EQM de l'estimateur de test préliminaire et la moins élevée des EQM de l'estimateur fondé sur les chiffres redressés et de l'estimateur fondé sur les chiffres non redressés. Lorsqu'il n'y a pas de biais, cette méthode produit une valeur optimale de c d'environ 1.88. Cela correspond à un CV critique (en valeur absolue) de 73% pour le niveau de sous-dénombrement estimé. L'EQM de cet estimateur est reproduite dans la figure 4.

2.2 Estimateur fondé sur les chiffres redressés du recensement

Dans ce cas, l'estimateur de U est \hat{U} . Son biais est UR et sa variance, σ^2 . Par conséquent, $\text{EQM}(\hat{U}) = \sigma^2 + UR^2$.

2.3 Estimateur de test préliminaire

En comparant les EQM des deux estimateurs précédents, on peut penser qu'il sera préférable d'utiliser les chiffres redressés, par opposition aux chiffres non redressés, dès que

$$(1) \quad \sigma^2 < U^2(1 - R^2).$$

Bien que les paramètres de cette inéquation soient inconnus, on peut les estimer (sauf R) à l'aide des études de mesure de la couverture, de là la possibilité de se servir de ces estimations pour construire un test statistique qui permettra de vérifier l'hypothèse que l'inégalité est juste. Le résultat du test détermine le choix de l'estimateur (d'où le nom d'estimateur de test préliminaire).

De façon précise, supposons que $|R| < 1$ (condition indispensable pour que (1) se vérifie) et que $U \sim N(U(1 + R), \sigma^2)$, où σ^2 est connu. Alors, U^2/σ^2 suit une distribution de $\chi^2_{(1)}$ non centrale avec comme paramètre de non-centralité $\lambda = U^2(1 + R)^2/2\sigma^2$. L'hypothèse nulle $H_0: \sigma^2 \geq U^2(1 - R^2)$ équivaut à l'hypothèse $H_0: \lambda \leq (1 + R)^2/2(1 - R)$. On pourrait donc, par exemple, procéder à un redressement si $U^2/\sigma^2 > c$, la valeur critique $c \geq 0$ étant choisie de manière que

$$(2) \quad \alpha = \Pr \left\{ \chi^2_{\left(1, \frac{1+R}{2(1-R)}\right)} \geq c \right\}.$$

où α est le niveau de signification du test. Ce test est une version particulière d'un test plus général proposé par Toro-Vizcarrondo et Wallace (1968).

Notons que U^2/σ^2 est l'inverse du carré du coefficient de variation (CV) estimé de \hat{U} . Cela revient à dire que pour effectuer un redressement, il faut un CV suffisamment petit (en valeur absolue).

En pratique, il faudrait remplacer R dans (2) par une estimation provisoire du biais relatif, disons r . Royce (1991) étudie la sensibilité de c par rapport à R dans le cas d'un test unilatéral (on a utilisé en l'occurrence une distribution normale au lieu d'une distribution de χ^2). Par exemple, étant donné un niveau de signification de 2,5%, on a observé que la réduction du CV critique ne dépassait pas six points de pourcentage environ (de 33,8 à 27,1%), même lorsque le biais relatif atteignait 50%.

Si σ n'est pas connu mais qu'il existe une estimation $\hat{\sigma}$, on peut construire un test semblable en supposant que

$$(3) \quad \frac{r\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(v)}$$

qui est indépendant de \hat{U} . On obtient alors un test fondé sur une distribution F non centrale. Pour obtenir plus de détails sur la construction de ces tests, le lecteur est prié de consulter Judge et Bock (1978). Afin de déterminer l'EQM de l'estimateur de test préliminaire, notons que cet estimateur peut s'écrire $U^P = IU$ où

le logement a été classé par erreur parmi les logements inoccupés. Du point de vue technique toutefois, la question étudiée se rapproche sensiblement de la question du redressement des chiffres du recensement qui retient l'attention de nombreux organismes statistiques depuis quelques années.

Les deux questions fondamentales qui se posent relativement au redressement sont de savoir I) dans quelle mesure un redressement améliore les chiffres du recensement et 2) quelles sont les meilleures méthodes de redressement. Dans cet article, nous comparons la précision de divers estimateurs d'une série de totaux de population en utilisant comme critère l'erreur quadratique moyenne pondérée (EQMP).

Dans la section 2, nous commençons par étudier le cas d'un total de population unique. Nous calculons puis comparons les erreurs quadratiques moyennes de quatre estimateurs possibles: estimateur fondé sur les chiffres non redressés du recensement, estimateur fondé sur les chiffres redressés du recensement, estimateur de test préliminaire et estimateur composé. Dans la section 3, nous étendons l'analyse à des totaux de population multiples et à des fonctions de totaux de population comme les proportions de population et les taux de croissance. La section 4 porte sur l'estimation pour petites régions; nous y étudions des méthodes d'estimation, et notamment l'estimation synthétique et un cas particulier de l'estimation synthétique appelée redressement général. Enfin, dans la section 5, nous proposons des sujets de recherche pour l'avenir.

L'élaboration des estimateurs présentés dans cet article repose sur deux hypothèses. Premièrement, les opérations de redressement doivent produire des estimations qui soient cohérentes entre les divers niveaux d'aggrégation géographique et les divers groupes démographiques et cohérentes d'une période à l'autre. Aux yeux des utilisateurs, il est essentiel que la somme des parties égale le tout et qu'il n'y ait pas de bris de continuité majeure dans les séries d'estimations. Deuxièmement, le redressement doit reposer sur les résultats combinés des deux études de mesure de la couverture de Statistique Canada, à savoir la contre-vérification des dossiers, qui sert à mesurer le sous-dénombrement brut, et l'étude du surdénombrement, qui sert à mesurer le surdénombrement brut. Les deux études sont exposées à l'erreur d'échantillonnage et à l'erreur non due à l'échantillonnage.

2. TOTAL DE POPULATION UNIQUE

Nous allons tout d'abord définir et comparer quatre estimateurs pour un total de population unique, l'erreur quadratique moyenne (EQM) servant de critère dans la comparaison.

Soient: Y l'effectif recensé;

T l'effectif réel (inconnu) de la population (à estimer);

U le sous-dénombrement net réel, c.-à-d. $U = T - Y$;

U une estimation de U tirée des études de mesure de la couverture;

σ^2 la variance de U ;

R le biais relatif de U , c.-à-d. $R = E(U)/U - 1$.

Pour les quatre estimateurs, on peut exprimer la valeur estimée de T comme la somme de l'effectif recensé et de la valeur estimée de U . Par conséquent, l'EQM de la valeur estimée de T sera la même que celle de la valeur estimée de U correspondant. Les EQM (et EQMP dans les sections subséquentes) sont calculées pour des répétitions hypothétiques des études de mesure de la couverture, où les chiffres du recensement sont considérés comme des quantités fixes.

2.1 Estimateur fondé sur les chiffres non redressés du recensement

Dans ce cas particulier, la valeur estimée de U est nulle. Son biais est égal à $-U$ et sa variance, égale à zéro. Par conséquent, $EQM(U^c) = U^2$.

Une comparaison d'estimateurs d'un ensemble de totaux de population

DON ROYCE¹

RÉSUMÉ

Les données du Programme des estimations démographiques (PED) de Statistique Canada sont toujours énoncées en fonction du recensement le plus récent sans qu'il soit tenu compte de l'erreur de couverture dans ce recensement. Or, depuis qu'ils ont constaté une forte hausse du niveau de sous-dénombrement lors du recensement de 1986, Statistique Canada étudie la possibilité de redresser l'effectif de la population de base du PED pour tenir compte du sous-dénombrement net. Dans cet article, on définit et compare quatre estimateurs de l'effectif de la population de base: les chiffres non redressés du recensement, les chiffres redressés, un estimateur de test préliminaire et un estimateur composite. L'élément de comparaison, en l'occurrence, est une généralisation des fonctions de risque proposées antérieurement connue sous le nom d'erreur quadratique moyenne pondérée (EQMP). L'EQMP s'applique non seulement aux totaux de population, mais aux fonctions de totaux de population, comme les proportions de population et les taux de croissance d'un recensement à l'autre. Il est aussi question de l'utilisation de l'EQMP dans l'élaboration et l'évaluation des estimateurs pour petites régions dans le contexte du redressement des chiffres du recensement.

MOTS CLÉS: Ajustement du recensement; sous-dénombrement; estimation pour les petites régions.

1. INTRODUCTION

Le Programme des estimations démographiques (PED) de Statistique Canada fournit une multitude de données détaillées sur les caractéristiques et la répartition de la population canadienne pour la période quinquennale qui sépare deux recensements. Les estimations intercen-sitaires de la population servent à de nombreuses opérations importantes comme le calcul des paiements de transfert – qui se chiffrent à plusieurs milliards de dollars – entre l'administration fédérale et les administrations provinciales, l'estimation de variables démographiques importantes comme le taux de natalité et le taux de mortalité, la détermination des niveaux d'immigration futurs et la pondération dans les enquêtes démographiques courantes comme l'enquête mensuelle sur la population active.

Le PED repose depuis toujours sur les données du recensement le plus récent sans que l'erreur de couverture n'ait jamais été prise en considération. En 1986 toutefois, on a observé une hausse appréciable du niveau de sous-dénombrement par rapport aux recensements précédents et ce niveau était toujours aussi variable d'une région à l'autre et d'un groupe démographique à l'autre. Cela a perturbé le PED et les nombreux autres programmes qui utilisent des estimations démographiques. C'est pourquoi on a mis sur pied au début de 1989 un projet qui visait à discuter de l'opportunité de redresser les estimations démographiques postcensitaires en fonction de l'erreur de couverture estimée dans le recensement et de la manière d'effectuer ce redressement. La recherche décrite dans cet article s'inscrivait dans le cadre de ce projet. Pour une description plus générale du projet, voir Royce (1992).

Il convient de souligner que ce redressement ne toucherait que les estimations démographi-ques. Les données du recensement de 1991 seront publiées sans qu'il y ait de redressement pour le sous-dénombrement, exception faite des redressements mineurs normalement effectués pour tenir compte du sous-dénombrement chez les résidents temporaires et chez les personnes dont

¹ Don Royce, chef de la Section de la qualité des données du recensement, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (avec discussion). *Statistical Science*, 1, 3-39.

HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.

HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.

HUANG, E.T., ISAKI, C.T., et TSAY, J.H. (1991). Modelling PES adjustment factors using 1988 dress rehearsal data. *Proceedings of the Social Statistics Section, American Statistical Association*, à paraître.

ISAKI, C.T., HUANG, E.T., et TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 post enumeration survey. *Proceedings of the Social Statistics Section, American Statistical Association*, à paraître.

KACKAR, R.N., et HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

MORRIS, C. (1983). Parametric Empirical Bayes inference and applications. *Journal of the American Statistical Association*, 78, 47-65.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2^{ème} édition). New York: Wiley.

On a donc:

$$E\left[-\frac{d^2\log L}{d(\sigma^2)^2}\right] = -1/2\operatorname{tr}(\Sigma^{-2}) + \operatorname{tr}(\Sigma^{-2}) = 1/2\operatorname{tr}(\Sigma^{-2}).$$

En approximant $E[(\hat{\sigma}^2 - \sigma^2)^2]$ par

$$E\left[-\frac{d^2\log L}{d(\sigma^2)^2}\right]^{-1},$$

ce qui se justifie par la théorie asymptotique du maximum de vraisemblance, on déduit de (A.16):

$$E[(\hat{\Theta}_{EB} - \Theta)(\hat{\Theta}_{EB} - \Theta)^T] \approx 2(\operatorname{tr}(\Sigma^{-2}) - {}^1VK^3V. \tag{A.22}$$

En combinant les formules (A.7) à (A.10) avec (A.22), on obtient:

$$\operatorname{MSE}(\hat{\Theta}_{EB}) \doteq V - V\Sigma^{-1}V + V\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V + VK^3V[2(\operatorname{tr}\Sigma^{-2}) - {}^1](\text{A.23})$$

Le remplacement de Σ par Σ fournit l'approximation donnée en (2.8), ce qui termine la preuve du théorème 2.

REMERCIEMENTS

Cet article fait suite à un travail qui a été soutenu financièrement par la subvention SES 87-13643 "Recherche sur le terrain pour améliorer la base de données gouvernementale en sciences humaines" de la National Science Foundation. La recherche a été menée au Bureau of the Census des E.-U. pendant que les auteurs participaient au programme de recherches conjoint de l'American Statistical Association et du Census Bureau, programme soutenu financièrement par le Census Bureau et par la subvention de la NSF. Les opinions, constatations, conclusions ou recommandations exprimées ici sont celles des auteurs et ne reflètent pas nécessairement les opinions de la National Science Foundation ou du Bureau of the Census.

BIBLIOGRAPHIE

CHILDEERS, D.R., et HOGAN, H. (1990). Results of the 1988 dress rehearsal post enumeration survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 547-552.

DATTA, G.S. (1990). Bayesian prediction in mixed linear models with applications in small area estimation. Thèse de doctorat non-publiée. University of Florida.

DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., et SCHULTZ, L.K. (1990). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress-Rehearsal data. Technical Report No. 376. Department of Statistics, University of Florida.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for Small Places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Puisque $g(\sigma^2) = Y - V \Sigma^{-1} [Y - X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y]$, on obtient par des techniques de différentiation matricielle:

$$\frac{dg}{d\sigma^2} = V[\Sigma^{-1} - \Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}] \Sigma^{-1} (Y - X\hat{\beta}). \quad (A.12)$$

$$E \left[\frac{dg}{d\sigma^2} \frac{d\sigma^2}{d\sigma^2} \right] = VK \Sigma^{-1} E[(Y - X\hat{\beta})(Y - X\hat{\beta})^T \Sigma^{-1} K^T V]. \quad (A.13)$$

Cependant, des calculs algébriques simples nous donnent:

$$E[(Y - X\hat{\beta})(Y - X\hat{\beta})^T] = \Sigma - X(X^T \Sigma^{-1} X)^{-1} X^T = \Sigma K \Sigma. \quad (A.14)$$

D'où on déduit de (A.13):

$$E \left[\frac{dg}{d\sigma^2} \frac{d\sigma^2}{d\sigma^2} \right] = VK^3 V. \quad (A.15)$$

Si on utilise encore une autre approximation, il s'ensuit de (A.11) et de (A.15) que:

$$E[(\hat{\theta}_{EB} - \theta)(\hat{\theta}_{EB} - \theta)^T] = E(\hat{\sigma}^2 - \sigma^2)^2 VK^3 V. \quad (A.16)$$

Pour estimer $E(\hat{\sigma}^2 - \sigma^2)^2 = \text{MSE}(\hat{\sigma}^2)$, nous procédons de la façon suivante. Puisque $Y \sim N(X\beta, \Sigma)$, exprimons la fonction de vraisemblance par:

$$L(\sigma^2) \propto |\Sigma|^{-1/2} \exp[-1/2(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)]. \quad (A.17)$$

D'où

$$\frac{d \log L}{d \sigma^2} = -1/2 \frac{d \log |\Sigma|}{d \sigma^2} - 1/2 \frac{d \log L}{d \sigma^2} = -1/2 \frac{d \log L}{d \sigma^2} \left[(Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \right]; \quad (A.18)$$

$$\frac{d^2 \log L}{d^2 \sigma^2} = -1/2 \frac{d^2 \log |\Sigma|}{d^2 \sigma^2} - 1/2 \frac{d^2 \log L}{d^2 \sigma^2} \left[(Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \right]. \quad (A.19)$$

Comme auparavant, notons d_1, \dots, d_m les valeurs propres de V .

Alors $\log |\Sigma| = \Sigma_m^{l=1} \log(\sigma^2 + d_l)$. D'où

$$\frac{d^2 \log L}{d^2 \sigma^2} \left[\Sigma \right] = -\Sigma_m^{l=1} (\sigma^2 + d_l)^{-2} = -\text{tr}(\Sigma^{-2}). \quad (A.20)$$

Si on utilise (A.20) et la différentiation matricielle, il découle de (A.19) que:

$$\frac{d^2 \log L}{d^2 \sigma^2} \left[\frac{d(\sigma^2)}{d\sigma^2} \right] = 1/2 \text{tr}(\Sigma^{-2}) - (Y - X\beta)^T \Sigma^{-3} (Y - X\beta). \quad (A.21)$$

ce qui se réduit après simplification à :

$$|V^{-1}| |I + \sigma^{-2}V| |X^T(I + \sigma^{-2}V)^{-1}X| \div |X^TX| \propto |I + \sigma^{-2}V| |X^T \Sigma^{-1}X|. \quad (\text{A.5})$$

De plus, après quelques calculs, il s'ensuit que :

$$V^{-1} - V^{-1}E^{-1}V^{-1} = F. \quad (\text{A.6})$$

La preuve de la partie (ii) du théorème découle maintenant de (A.4) à (A.6) et de la remarque que $f(\sigma^2 | y) \propto f(\sigma^2, y)$. Notons cependant que la distribution *a posteriori* de σ^2 pour $X = y$ fixé est propre.

Preuve du théorème 2. Encore une fois, nous ne ferons qu'esquisser la preuve. On peut trouver les détails dans Datia et coll. (1991).

Rappelons que

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y.$$

Posons

$$\hat{\theta} = X\hat{\beta} + \sigma^2 \Sigma^{-1} (Y - X\hat{\beta}).$$

Observons maintenant que (i) $\hat{\theta}$ est le meilleur prédicteur sans biais de θ (à cause de la normalité) pour chaque σ^2 fixé et (ii) $E(\hat{\theta}_{EB} - \theta) = 0$, étant donné que $\hat{\sigma}^2$ est l'estimateur linéaire moyen de σ^2 (voir Kackar et Harville (1984)). En utilisant maintenant le lemme 3.3.1 de Datia (1990), on voit que $\hat{\theta}_{EB} - \theta$ et $\hat{\theta}$ sont indépendants. Donc on a :

$$E[(\hat{\theta}_{EB} - \theta)(\hat{\theta}_{EB} - \theta)^T] =$$

$$E[(\hat{\theta}_{EB} - \theta)(\hat{\theta}_{EB} - \theta)^T] + E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]. \quad (\text{A.7})$$

Posons ensuite $\hat{\theta}^B = X\hat{\beta} + \sigma^2 \Sigma^{-1} (Y - X\hat{\beta})$. Un raisonnement standard nous donne :

$$E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = E[(\hat{\theta} - \hat{\theta}^B)(\hat{\theta} - \hat{\theta}^B)^T] + E[(\hat{\theta}^B - \theta)(\hat{\theta}^B - \theta)^T]. \quad (\text{A.8})$$

Nous pouvons déduire de nos calculs précédents que :

$$E[(\hat{\theta}^B - \theta)(\hat{\theta}^B - \theta)^T] = V - V\Sigma^{-1}V. \quad (\text{A.9})$$

De plus,

$$E[(\hat{\theta} - \hat{\theta}^B)(\hat{\theta} - \hat{\theta}^B)^T] = V\Sigma^{-1}X(X^T \Sigma^{-1}X)^{-1}X^T \Sigma^{-1}V. \quad (\text{A.10})$$

Enfin, posons $\hat{\theta} = g(\hat{\sigma}^2)$ et $\hat{\theta}_{EB} = g(\hat{\sigma}^2)$. En utilisant l'approximation de Taylor du premier degré, on obtient :

$$E[(\hat{\theta}_{EB} - \hat{\theta})(\hat{\theta}_{EB} - \hat{\theta})^T] \approx E\left[(\hat{\sigma}^2 - \sigma^2)^2 \frac{dg(\sigma^2)}{d\sigma^2} \frac{dg(\sigma^2)}{d\sigma^2}\right]. \quad (\text{A.11})$$

régionale totale obtenues après application des facteurs de redressement bruts. Les facteurs de redressement lissés et ajustés de façon proportionnelle ont ensuite été appliqués aux chiffres de population correspondants du recensement au niveau des îlots de recensement. Puis les résultats ont été arrondis par groupes d'îlots à l'entier le plus proche, de façon que chaque cellule d'un îlot soit arrondie vers le haut ou vers le bas pour donner un entier et que les sommes de contrôle ne changent pas par plus d'une personne.

Les méthodes utilisées pour redresser les chiffres du recensement de 1990 étaient précisées à l'avance et toute l'opération devait être menée selon un échéancier très serré. Le Bureau of the Census a recommandé d'utiliser les chiffres redressés du recensement de 1990. Des avis partagés à parts égales ont été exprimés sur la question par une commission spéciale, dont les membres avaient été choisis par le secrétaire au Commerce. Après avoir évalué les éléments mis en preuve de part et d'autre, le secrétaire a décidé de ne pas utiliser les chiffres redressés. La question est maintenant devant les tribunaux. Un sujet d'actualité est l'utilisation possible des chiffres redressés pour l'estimation post-censitaire. Des recherches sont actuellement menées pour en arriver à de meilleurs chiffres redressés pour l'estimation post-censitaire.

ANNEXE - PREUVES DES THÉORÈMES

Preuve du théorème 1. Nous ne ferons qu'esquisser la preuve. On peut trouver les détails dans Datta et coll. (1991). La fdp (impropre) à quatre variables de Y , θ , β et σ^2 est donnée par:

$$f(y, \theta, \beta, \sigma^2) \propto \exp \left[-1/2 (y - \theta)_T V^{-1} (y - \theta) \right] \sigma^{-m} \exp \left[-1/(2\sigma^2) \|\theta - X\beta\|_2^2 \right], \quad (\text{A.1})$$

où $\|\cdot\|$ dénote la norme euclidienne. Si on écrit $P_X = X(X^T X)^{-1} X^T$, $\|\theta - X\beta\|_2^2$, on a

$$\|\beta - (X^T X)^{-1} X^T \theta\|_2^2 = (X^T X)^{-1} X^T \theta^T \theta + \theta^T (I - P_X) \theta,$$

Si on intègre maintenant par rapport à β dans (A.1), il s'ensuit que la fdp impropre à trois variables de Y , θ et σ^2 est:

$$f(y, \theta, \sigma^2) \propto \sigma^{-(m-p)} \exp \left[-1/2 (y - \theta)_T V^{-1} (y - \theta) - 1/(2\sigma^2) \theta^T (I - P_X) \theta \right]. \quad (\text{A.2})$$

Si on écrit ensuite $E = V^{-1} + \sigma^{-2} (I - P_X)$, il s'ensuit après quelques simplifications que:

$$(y - \theta)_T V^{-1} (y - \theta) + \sigma^{-2} \theta^T (I - P_X) \theta =$$

$$(\theta - E^{-1} V^{-1} y)_T E (\theta - E^{-1} V^{-1} y) + y_T (V^{-1} - V^{-1} E^{-1} V^{-1}) y. \quad (\text{A.3})$$

Par conséquent, la distribution *a posteriori* de θ pour σ^2 donné et $Y = y$ est $N(E^{-1} V^{-1} y, E^{-1})$. En utilisant la formule bien connue pour l'inversion d'une matrice $(A + BDB^T)^{-1} = A^{-1} - A^{-1} B (D^{-1} + B^T A^{-1} B)^{-1} B^T A^{-1}$ (voir par exemple l'exercice 2.9, p. 33 de Rao (1973)), on obtient $E^{-1} = G$, ce qui complète la preuve de la première partie du théorème. En se servant ensuite de (A.3) et en intégrant par rapport à θ dans (A.2), on obtient la fdp (impropre) à deux variables de Y et σ^2 :

$$f(y, \sigma^2) \propto \sigma^{-(m-p)} |E|^{-1/2} \exp \left[- (1/2) y_T (V^{-1} - V^{-1} E^{-1} V^{-1}) y \right]. \quad (\text{A.4})$$

Par l'exercice 2.4, p. 32 de Rao (1973), il s'ensuit que

$$|E| = |V^{-1} + \sigma^2 (I - P_X)| = |\Delta| \div |\sigma^2 X^T X|$$

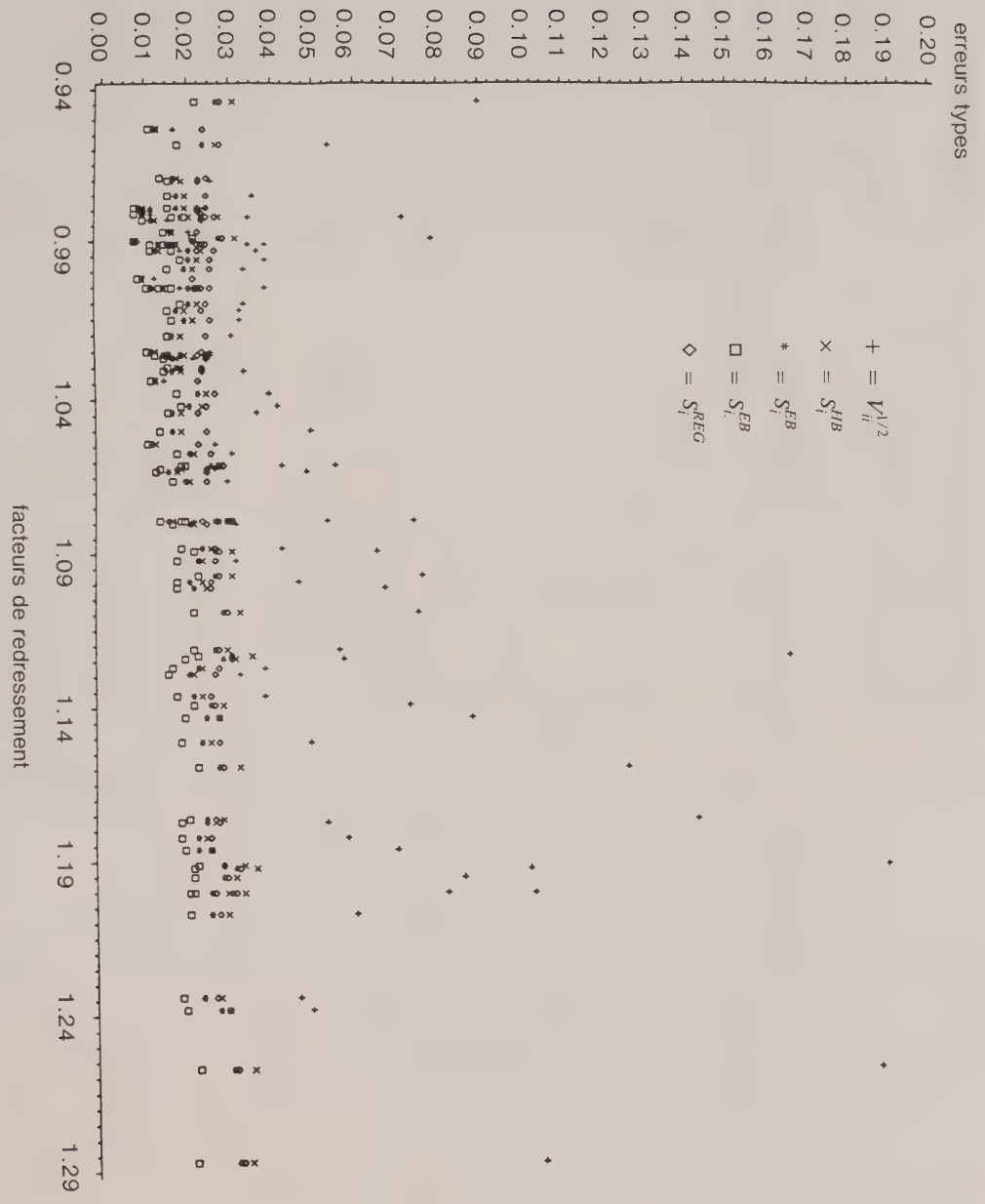


Figure 3. ET des facteurs de redressement par facteurs de redressement.

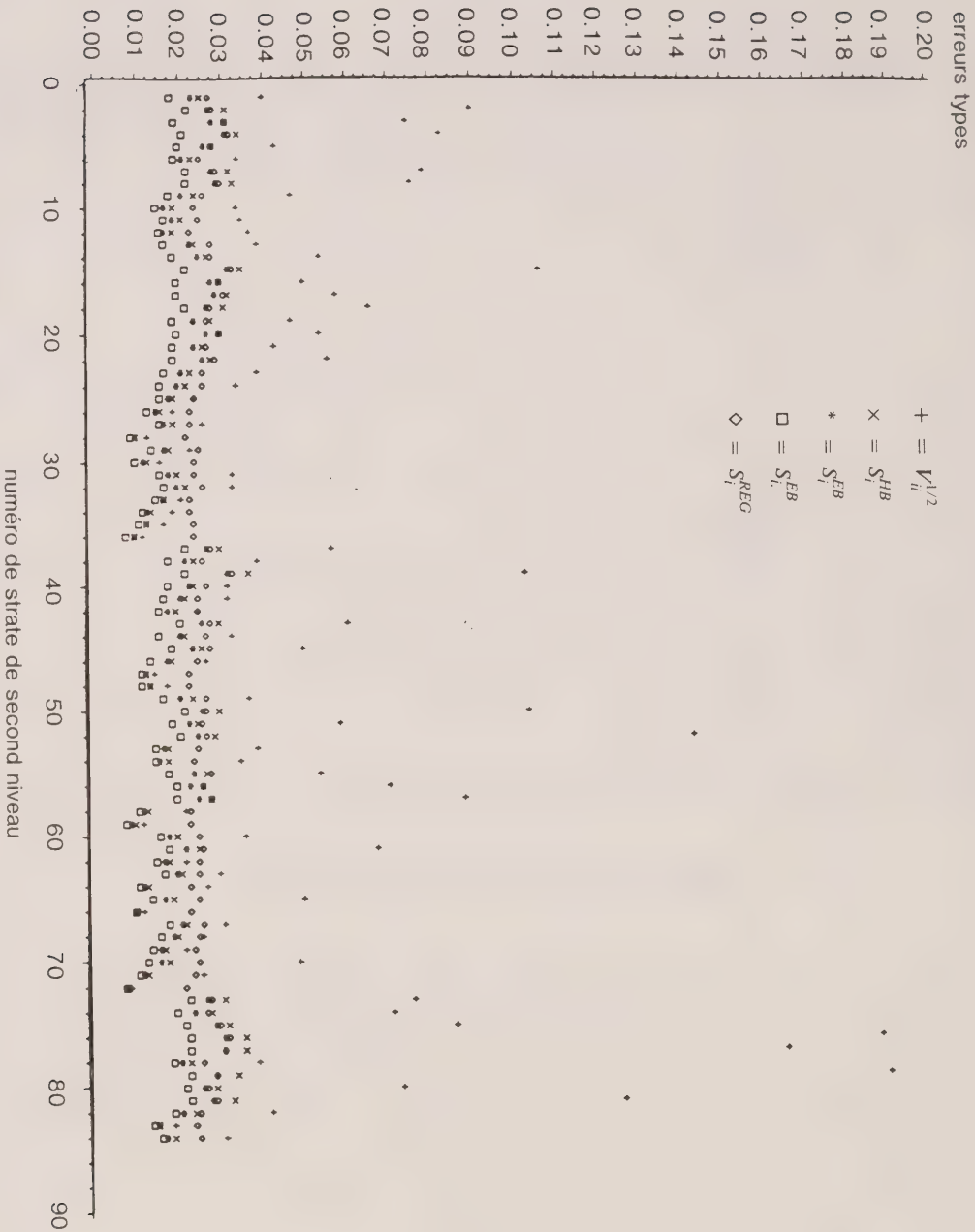


Figure 2. ET des facteurs de redressement par strate de second niveau.

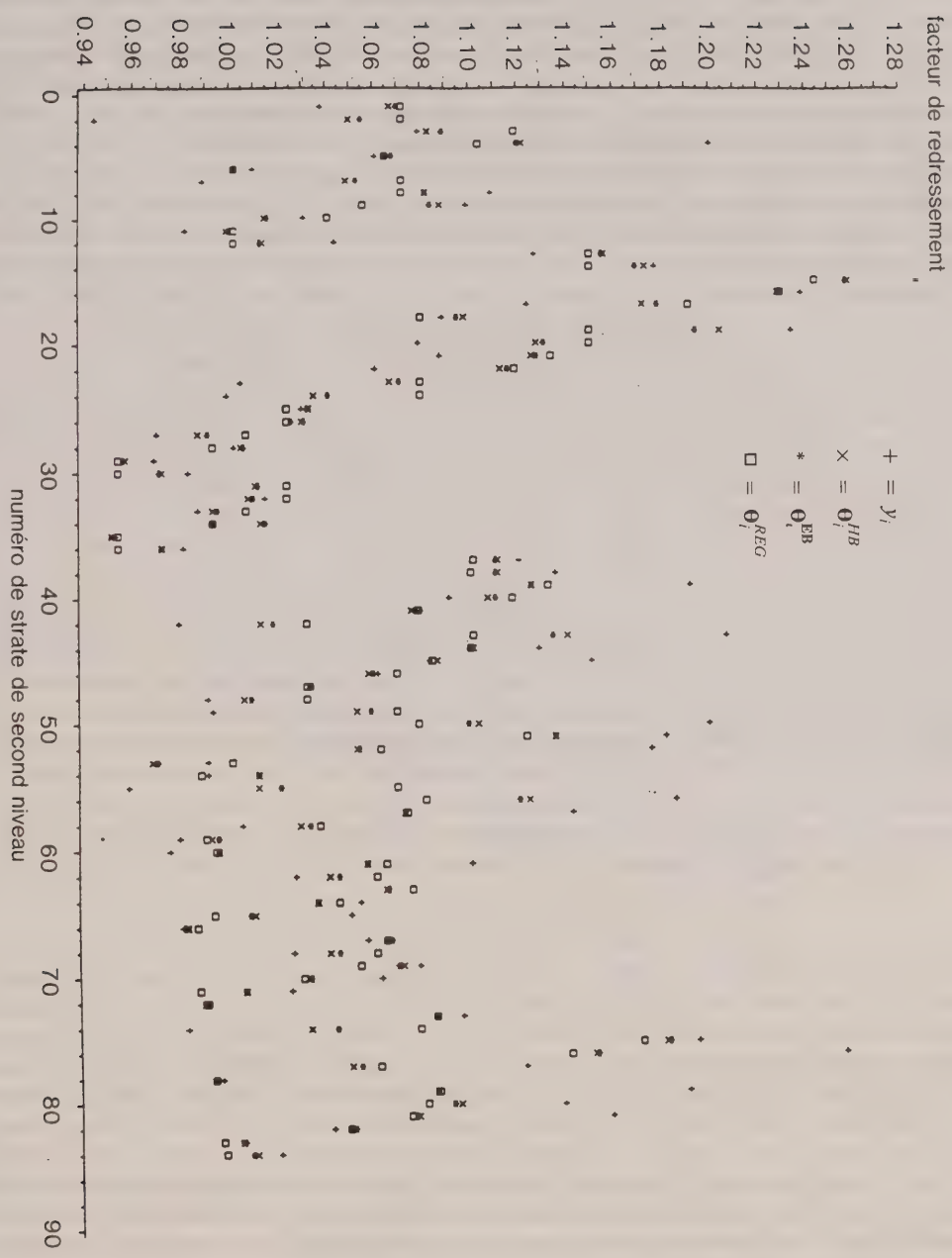


Figure 1. Facteurs de redressement par strate de second niveau.

Les figures 1 et 2 représentent graphiquement les facteurs de redressement estimés et les erreurs types, par strate de second niveau. Les 12 premières strates de second niveau se rapportent aux non-propriétaires blancs non hispaniques de Saint Louis, les strates 13 à 24 se rapportent à tous les autres non-propriétaires de Saint Louis, les strates 25 à 36 se rapportent à tous les autres propriétaires blancs non hispaniques de Saint Louis et les strates 37 à 48 se rapportent à tous les autres propriétaires de Saint Louis. Les strates 49 à 60 se rapportent aux Blancs non hispaniques vivant dans des régions avec registre d'adresses informatisé (RAI) du centre-est du Missouri, les strates 61 à 72 se rapportent aux personnes blanches non hispaniques vivant dans des régions sans RAI du centre-est du Missouri et les strates 73 à 84 se rapportent à toutes les autres personnes du centre-est du Missouri.

Dans chaque groupe de 12 strates de second niveau, les six premières se rapportent aux personnes de sexe masculin dans les groupes d'âge 0-9, 10-19, 20-29, 30-44, 45-64 et 65+. Un examen de la figure 1 permet de constater que les facteurs de redressement bruts du groupe des Blancs non hispaniques ont tendance à être inférieurs à ceux de l'autre groupe, sauf pour la région avec RAI du centre-est du Missouri. On voit à la figure 2 que la même observation est presque vérifiée dans le cas des erreurs types brutes. La figure 3 présente un graphique des erreurs types estimées en fonction des facteurs de redressement.

Les figures 1 à 3 nous amènent à tirer plusieurs conclusions intéressantes.

(1) Dans chaque strate, les erreurs types estimées des estimateurs HB et EB des facteurs de redressement sont beaucoup plus petites que les erreurs types des facteurs de redressement bruts lorsqu'on les compare aux ETD non redressées.

(2) Les estimateurs EB sont supérieurs aux estimateurs de régression dans chacune des 84 strates en ce qu'ils permettent une réduction de l'erreur type estimée. Bien que les estimateurs HB ne soient pas supérieurs aux estimateurs de régression dans toutes les strates, l'amélioration est considérable dans la plupart des strates.

(3) La représentation graphique des données montre que l'écart entre les estimations ponctuelles $\hat{\theta}_{EB}^i$ et $\hat{\theta}_{HB}^i$ est assez petit. En fait, la différence est toujours inférieure à 1% (et souvent de beaucoup).

(4) Les erreurs types *a posteriori* associées aux estimations HB (s_{HB}^i) sont toujours supérieures aux EQM approximatives des estimations EB (s_{EB}^i). Comme nous l'avons déjà mentionné, les deux erreurs types ne sont pas forcément égales. Il nous semble que les erreurs types approximatives des estimations EB sont souvent de légères sous-estimations. Cependant, une comparaison entre s_{EB}^i et s_{HB}^i révèle qu'une méthode EB de type naïf (avec erreurs types associées s_{EB}^i) peut grandement sous-estimer les erreurs types estimées en ne tenant pas compte de l'incertitude liée à l'estimation de σ^2 . Ce défaut est en grande partie rectifié par s_{EB}^i , qui est fondé sur des approximations du deuxième ordre.

Au moment où nous avons révisé cet article, le redressement du recensement décennal de 1990 était terminé. Ce redressement a été effectué par la méthode de l'estimation EB. En gros, pour modéliser les facteurs de redressement, on a utilisé la plupart des étapes suivies au moment de la répétition générale de 1988. Il y avait toutefois plusieurs différences. Dans le redressement de 1990, les facteurs de redressement ont été modélisés pour chacune des quatre régions de recensement, avec un ensemble spécial de facteurs de redressement pour les réserves indiennes. Le nombre de facteurs de redressement variait de 12 pour les données se rapportant aux Indiens à 456 dans une des régions. De plus, les variannes estimées des facteurs de redressement bruts ont été lissées à l'aide de modèles de régression. Le lissage des variannes estimées avait tendance à réduire les grandes variannes estimées et à augmenter les petites. L'effet net a été une augmentation de la contribution aux estimations EB des facteurs de redressement associés qui ont une grande variance estimée, et vice versa. D'autres différences concernent l'emploi de procédés de détection des observations aberrantes pour le lissage des variannes comme pour celui des facteurs de redressement. Enfin, les estimations EB pour les strates de second niveau ont été ajustées de façon proportionnelle aux estimations de la population

Un estimateur naïf de la matrice des variances $\hat{\Theta}^{EB}$ est $V - V\hat{\Sigma}^{-1}V$. Cette matrice sous-estimée de façon flagrante la matrice des variances puisqu'elle ne tient pas compte de l'incertitude liée à l'estimation de β et σ^2 . Si on suppose σ^2 connu, et si on attribue à β une distribution *a priori* uniforme sur R^p ($m \geq p + 3$), alors l'estimateur HB de Θ est le même que celui de $\hat{\Theta}^{BLUP}$ et la matrice des variances *a posteriori* est alors donnée par $M = V - V\hat{\Sigma}^{-1}V + V\hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}V$. On en déduit de façon immédiate que $E[(\hat{\Theta}^{BLUP} - \Theta)(\hat{\Theta}^{BLUP} - \Theta)^T] = M$, où l'espérance est calculée sur la distribution à deux variables de Y et de Θ , donnée en I et II. Dans le langage bayésien, on peut donc interpréter $V\hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}V$ comme le supplément de variabilité *a posteriori* causé par l'incertitude sur β , alors qu'en terminologie classique le même phénomène peut s'interpréter comme le supplément d'EQM causé par la même incertitude.

Le traitement d'un σ^2 inconnu est un autre problème à résoudre. La méthode de Bayes nous permet de trouver la distribution *a posteriori* de σ^2 pour $Y = y$, alors que, même sans introduire une estimation *a priori* de Θ , il demeure possible de trouver une approximation de l'EQM de $\hat{\Theta}^{EB}$ en modifiant un argument de Kackar et Harville (1984) ou de Prasad et Rao (1990).

Le théorème pertinent, dont la preuve est reportée en annexe, est énoncé ci-dessous.

Théorème 2. On peut estimer de façon approximative l'EQM de $\hat{\Theta}^{EB}$ par :

$$(2.8) \quad MSE(\hat{\Theta}^{EB}) \approx V - VKV + (VK^3V) [2(u\hat{\Sigma}^{-2})^{-1}],$$

$$(2.9) \quad K = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}.$$

où

Le troisième terme du membre de droite de (2.8) peut être interprété comme le supplément d'erreur quadratique moyenne dû à l'incertitude de l'estimation de σ^2 . Harville (1985) donne une décomposition générale de l'erreur de prévision.

Bien que les variances *a posteriori* $V(\Theta | y)$ associées à l'estimateur HB $\hat{\Theta}^{HB}$ de Θ et l'EQM estimée de l'estimateur $\hat{\Theta}^{EB}$ de Θ proviennent de théories inspirées de deux philosophies différentes de l'inférence, elles ont en commun d'essayer de représenter l'incertitude causée par l'estimation de la variance du modèle. Pour mieux comprendre cette similitude, il faut remarquer que, si on écrit $K = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}$, alors on a :

$$(2.10) \quad E[V(\Theta | \sigma^2, y)] = G = V - VKV$$

et $E(G | y)$ est approximée par $V - VKV$, qui est un des deux termes de (2.8). On a aussi $(\Theta | \sigma^2, y) = GV^{-1}y$, et on peut démontrer après quelques simplifications que $GV^{-1} = I - VK$. On a donc $V(GV^{-1}y | y) = VV(K | y)V$ et $V(K | y)$ est apparemment approximé par $K^3[2(tr\hat{\Sigma}^{-2})^{-1}]$. Cependant, l'approximation par EQM de $\hat{\Theta}^{EB}$ ne correspond pas toujours parfaitement à $V(\Theta | y)$, comme le prouvent les calculs numériques de la section 3. Une des hypothèses d'Ericksen et Kadane (1985) est celle du σ^2 connu. Freedman et Navidi (1986) ont insisté sur la nécessité d'une estimation de σ^2 et, dans les théorèmes 1 et 2, nous avons rendu compte de cette source d'incertitude, tant de façon bayésienne que de façon fréquentiste. Notons que, au contraire de travaux antérieurs qui ne traitaient que de l'estimation du sous-dénombrément net de la population totale au niveau d'une ville et à celui du reste de l'Etat, nous nous intéressons à l'estimation de facteurs de redressement à des niveaux d'analyse plus fins. En termes de coûts d'exploitation, le redressement à des niveaux plus fins permet des économies considérables de temps et de coût de calcul par ordinateur, puisque les dossiers de recensement n'ont besoin d'être utilisés qu'une seule fois. Les modèles de redressement qui utilisent des niveaux géographiques plus élevés que les nôtres exigent plusieurs passages en machine des données du recensement parce qu'ils exigent une méthode pour distribuer le

En utilisant (2.2) et (2.3), on déduit $E(\Theta | y)$ et $V(\Theta | y)$ de (2.4) et (2.5) par intégration numérique. On peut simplifier quelque peu les calculs requis pour (2.1) à (2.3) en appliquant à V le théorème de décomposition spectrale. Ainsi on aura $V = PDP^T$, où $D = \text{Diag}(d_1, \dots, d_m)$, les d_i étant les valeurs propres de V , et $P = (\xi_1, \dots, \xi_m)$, les ξ_i étant les vecteurs propres orthornormaux correspondants. En utilisant l'orthogonalité de P , on obtient alors:

$$|\Sigma| = |\sigma^2 I + D| = \prod_{i=1}^m (\sigma^2 + d_i);$$

$$\Sigma^{-1} = P(\sigma^2 I + D)^{-1} P^T;$$

$$X^T \Sigma^{-1} X = (P^T X)^T (\sigma^2 I + D)^{-1} (P^T X);$$

$$F = P(\sigma^2 I + D)^{-1} P^T - P(\sigma^2 I + D)^{-1} (P^T X) \times$$

$$[(P^T X)^T (\sigma^2 I + D)^{-1} (P^T X)] (\sigma^2 I + D)^{-1}.$$

En pratique, l'intégration numérique sur σ^2 , qui demande qu'on évalue l'intégrande pour différentes valeurs de σ^2 , devient maintenant un peu plus simple puisque P et X sont connus et que $\sigma^2 I + D$ est une matrice diagonale. Examinons maintenant l'estimation EB. Dans ce cas, on n'utilise pas l'hypothèse III. En supposant β et σ^2 connus, on déduit d'abord de I et de II un estimateur de Bayes, à savoir la moyenne *a posteriori* de Θ . Cet estimateur est donné par:

$$\Theta_B = E(\Theta | Y, \beta, \sigma^2)$$

$$= (V^{-1} + \sigma^{-2} I)^{-1} (V^{-1} Y + \sigma^{-2} X \beta)$$

$$= \Sigma^{-1} (\sigma^2 Y + V X \beta). \quad (2.6)$$

La variance *a posteriori* correspondante est donnée par:

$$V(\Theta | Y, \beta, \sigma^2) = (V^{-1} + \sigma^{-2} I)^{-1} = V - V \Sigma^{-1} V.$$

Cependant, en pratique, β et σ^2 sont inconnus et sont évalués par la méthode du maximum de vraisemblance à partir de la distribution marginale de Y , qui est $N(X\beta, \Sigma)$. Ces estimateurs de vraisemblance maximale (EVM) sont dénotés par $\hat{\beta}$ et $\hat{\sigma}^2$, où $\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$, $\hat{\Sigma} = V + \hat{\sigma}^2 I$. En substituant dans (2.6) ces estimateurs de Σ , σ^2 et β , on obtient l'estimateur EB de Θ suivant:

$$\Theta_{EB} = \hat{\Sigma}^{-1} (\hat{\sigma}^2 Y + V X \hat{\beta}) = X \hat{\beta} + \hat{\sigma}^2 \hat{\Sigma}^{-1} (Y - X \hat{\beta}). \quad (2.7)$$

On peut aussi voir que l'estimateur donné par (2.7) est un meilleur prédicteur linéaire sans biais estimé (MPLSE). Supposons d'abord que σ^2 est connu et trouvons le MPLS $\Theta_{BLUP} = X \hat{\beta} + \sigma^2 \Sigma^{-1} (Y - X \hat{\beta})$ de Θ , où $\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$. Estimons ensuite σ^2 par $\hat{\sigma}^2$, son estimateur linéaire moyen (ELM), et de façon correspondante estimons $\hat{\Sigma}$ par $\hat{\Sigma}$. En remplaçant σ^2 et Σ par $\hat{\sigma}^2$ et $\hat{\Sigma}$ dans l'expression de Θ_{BLUP} , on obtient le MPLSE Θ_{EB} .

possible, dans le présent contexte, d'éviter certaines (mais non la totalité) des critiques lancées contre la méthode d'Ericsen-Kadane (1985) par Freedman et Navidi (1986). Dans la section 3, nous analysons les données réelles. Nous donnons les estimations empiriques, les estimations HB, les estimations EB et les estimations de régression des facteurs de redressement, ainsi que les erreurs types associées. La méthode HB et les méthodes EB qui tiennent compte de l'incertitude reliée aux paramètres *a priori* inconnus ont une efficacité équivalente. Par leur meilleure capacité de réduire les erreurs types estimées, elles sont toutes deux nettement supérieures aux estimations brutes comme aux estimations de régression. Enfin, une annexe contient certains des détails techniques du présent article.

2. ESTIMATION HB ET ESTIMATION EB

La présente section décrit les méthodes générales de l'estimation HB et EB pour certains modèles hiérarchiques. L'application de ces méthodes à l'estimation de facteurs de redressement est examinée dans la section 3.

Considérons le modèle hiérarchique suivant:

- I. $Y \mid \theta, \beta, \sigma^2 \sim N(\theta, V)$, où V est une matrice $m \times m$ connue définie positive;
- II. $\theta \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$;
- III. β et σ^2 sont marginalement indépendants, avec β suivant une loi uniforme sur (R^p) et σ^2 une loi uniforme sur $(0, \infty)$.

L'analyse HB est fondée sur les points I à III. En l'absence d'information *a priori* précise sur β et σ^2 , nous préférons utiliser des distributions *a priori* diffusées dans III. Nous avons aussi analysé les données en supposant la fonction de densité de probabilité (*fdp*) *a priori* de σ^2 proportionnelle à σ^{-2} sur $(0, \infty)$. Les résultats étaient assez semblables et nous n'en faisons pas état ici. Le théorème suivant a été démontré.

Théorème 1. Soit le modèle défini par (I), (II) et (III). Écrivons $\Sigma = V + \sigma^2 I$. Supposons que $m \geq p + 3$. Alors (i) la fdp conditionnelle de θ pour σ^2 donné et $Y = y$ est $N(GV^{-1}y, G)$, où

$$G = V - V\Sigma^{-1}V + V\Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}V; \quad (2.1)$$

(iii) la fdp conditionnelle de σ^2 pour $Y = y$ est

$$f(\sigma^2 \mid y) \propto |\Sigma|^{-1/2} |X^T\Sigma^{-1}X|^{-1/2} \exp(-1/2 y^T F y), \quad (2.2)$$

où

$$F = \Sigma^{-1} - \Sigma^{-1}X[X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}. \quad (2.3)$$

La démonstration du théorème est reportée à l'annexe. En utilisant les formules de l'espérance et de la variance conditionnelles, on obtient:

$$E(\theta \mid y) = E[E(\theta \mid \sigma^2, y) \mid y] = (E(GV^{-1} \mid y)) y; \quad (2.4)$$

$$V(\theta \mid y) = V[E(\theta \mid \sigma^2, y) \mid y] + E[V(\theta \mid \sigma^2, y) \mid y] = V(GV^{-1} \mid y) + E(G \mid y). \quad (2.5)$$

le redressement des chiffres du recensement en partant d'un modèle. Ils ont préconisé l'ajustement des facteurs de redressement calculés à partir des données de l'EP à un modèle de régression convenable. Cette approche, dont la justification apparaît dans Erikssen et Kadane (1985), est semblable à celle envisagée dans Fay et Herriot (1979) ou Morris (1983). Malgré les critiques adressées à l'approche Erikssen-Kadane par certains statisticiens (les plus sévères étant celles de Freedman et Navidi (1986)), la plupart des gens reconnaissent l'importance de l'approche basée sur un modèle en ce qui concerne le redressement. D'ailleurs, dans le présent article, à part quelques différences dans les hypothèses, que nous indiquerons plus loin dans la section 2, nous utilisons le modèle de Fay-Herriot ou d'Erikssen-Kadane pour analyser les données de la répétition générale de 1988 au Missourï. Cressie (1989) présente une approche basée sur un modèle différent, qui ne comporte pas de variables corrélées supplémentaires. Childers et Hogan (1990) contient une bonne description de l'EP menée dans le cadre de la répétition générale de 1988 au Missourï. Hogan et Wolter (1988) examinent les catégories d'erreur qui se produisent dans une EP et analysent un moyen pour les évaluer. En deux mots, le plan de sondage d'une EP est constitué d'un échantillon d'îlots, stratifié à un seul degré, et d'une estimation de système dual du nombre de personnes dans chaque strate de second niveau.

Le présent article commence au moment où l'enquête post-censitaire nous a permis de déterminer un ensemble de facteurs de redressement bruts estimés, ainsi que leurs covariances, ces éléments pouvant maintenant être utilisés pour construire un modèle fondé sur les données de la répétition générale du recensement effectuée en 1988 dans des régions d'essai au Missourï. Nous supposons aussi que nous disposons d'un ensemble de variables explicatives possibles définies au second niveau, qui peuvent être utilisées pour des régressions. Deux régions géographiques sont étudiées ici: un grand noyau urbain, la ville de Saint Louis, et un ensemble de régions de population moyenne, le centre-est du Missourï. Pour définir les strates de second niveau à Saint Louis, nous avons classé les personnes selon les catégories démographiques suivantes: (i) la race (blanche non hispanique, autres), (ii) propriétaires, non-propriétaires (locataires) de logements, (iii) le sexe (masculin, féminin); les groupes d'âge (0-9, 10-19, 20-29, 30-44, 45-64, 65+). Ce classement conduit à un total de $2 \times 2 \times 2 \times 6 = 48$ facteurs de redressement pour Saint Louis. Pour le centre-est du Missourï, les catégories de sexe et les groupes d'âge sont demeurés les mêmes qu'à Saint Louis, mais une nouvelle catégorie (i)' a remplacé les catégories (i) et (ii). Les individus y sont classés comme (a) Blancs non hispaniques vivant dans une région avec registre d'adresses informatisé (RAI), (b) Blancs non hispaniques vivant dans une région sans RAI et (c) autres dans toutes les régions. Pour le centre-est du Missourï, un total de $3 \times 2 \times 6 = 36$ facteurs de redressement ont été calculés, ce qui fait un total de 84 facteurs de redressement pour l'ensemble du modèle. Dans chaque région, les facteurs de redressement estimés étaient corrélés, à cause de l'utilisation d'un plan de sondage par grappes de blocs. Par conséquent, les matrices des covariances empiriques correspondant à Saint Louis et au centre-est du Missourï, de dimensions respectives 48×48 et 36×36 , étaient diagonales par blocs.

Dans la section 2 du présent article, nous décrivons une méthode générale fondée sur un modèle pour les facteurs de redressement lissés et les erreurs types associées. On utilise aussi bien la méthode hiérarchique de Bayes que la méthode empirique de Bayes. On peut aussi considérer la méthode EB comme une méthode d'analyse des composantes de la variance (voir par exemple Harville 1985). Nous donnons aussi la formule pour calculer les erreurs types *a posteriori* associées aux estimateurs HB. Soulignons ici que l'utilisation naïve d'une méthode EB peut conduire à de graves sous-estimations des erreurs types associées, étant donné qu'une méthode EB naïve ne tient pas compte de l'incertitude reliée à l'estimation des composantes inconnues de la variance. Cependant, Kackar et Harville (1984) et Prasad et Rao (1990) ont proposé des approximations intéressantes des erreurs quadratiques moyennes (EQM) des estimateurs EB. En suivant leur principe, nous avons déduit des formules pour obtenir les EQM estimées dans notre contexte. Nous indiquons aussi dans cette section de quelle façon il est

Méthode hiérarchique de Bayes et méthode empirique de Bayes pour le redressement du sous-dénombrement: données de la "répétition générale" du recensement, effectuée en 1988 au Missouri

G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI,
L.K. SCHULTZ et J.H. TSAY¹

RÉSUMÉ

Le présent article analyse une approche basée sur un modèle pour le redressement des données de la répétition générale du recensement de 1990, effectuée en 1988. Ces données ont été recueillies dans des régions d'essai au Missouri. L'objectif premier est d'élaborer des méthodes qui peuvent être utilisées pour modéliser les données de l'enquête post-censitaire d'avril 1991, qui a fait suite au recensement de 1990, et pour lisser les estimations des facteurs de redressement tirés de l'enquête. Nous proposons dans le présent article une méthode hiérarchique de Bayes (HB) et une méthode empirique de Bayes (EB) qui satisfont à cet objectif. Les estimateurs qui résultent de ces deux méthodes semblent permettre d'améliorer de façon constante les estimations basées sur un système de double collecte et les estimateurs de régression lissés.

MOTS CLÉS: Enquête post-censitaire; facteurs de redressement; estimation de système dual; méthode hiérarchique de Bayes; méthode empirique de Bayes; composantes de la variance; MPLSE; estimations de régression; erreurs types.

1. INTRODUCTION

Le présent article analyse une approche basée sur un modèle pour le redressement des données recueillies en 1988 dans des régions d'essai au Missouri pendant la "répétition générale" du recensement de 1990. L'objectif principal qui sous-tend cet exercice est d'élaborer des méthodes qui peuvent être utilisées pour modéliser les données de l'enquête post-censitaire (EP) d'avril 1991, enquête qui a fait suite au recensement de 1990, et de lisser les estimations d'enquête de ce qu'on appelle les "facteurs de redressement bruts". Ces facteurs de redressement bruts sont des rapports entre des estimations de la population totale (inconnue) et les chiffres correspondants du recensement de 1990. Ils sont calculés à différents niveaux d'agrégation (régions géographiques telles que villes, banlieues, etc.) et mis en regard de différentes catégories démographiques (telles que l'âge, le sexe, la race, etc.). Les catégories à double entrée sont appelées strates de second niveau.

Avant de poursuivre, il convient de donner un bref historique. Le redressement des chiffres du recensement décennal de 1980 aux États-Unis a donné lieu à un vif débat, qui dure depuis près de dix ans. Malgré les efforts intenses et soutenus et les dépenses énormes encourues par le Bureau of the Census des E.-U. pour réaliser une couverture quasi-complète au cours du recensement de 1980, de nombreux procès ont été intentés contre le Bureau par des États ou des villes pour exiger une révision des chiffres publiés. Dans un de ces litiges, dont le monde statistique a beaucoup parlé à la suite des articles d'Erickson et Kadane (1985) et Freedman et Navidi (1986), la ville de New York, entre autres, a poursuivi le Bureau of the Census. De nombreux statisticiens réputés ont comparu comme témoins experts d'un côté comme de l'autre. En particulier, Erickson et Kadane ont comparu pour le plaignant et ont proposé d'aborder

¹ G.S. Datta, University of Georgia, Athens, GA 30602; M. Ghosh, University of Florida, Gainesville, FL 32611; E.T. Huang, C.T. Isaki, L.K. Schultz et J.H. Tsay, U.S. Bureau of the Census, Washington, DC 20233.

- ERICSEN, E.P., KADANE, J.B., et TURKEY, J.W. (1989). Adjusting the 1980 Census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.
- FAY, R.E., III, et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FELLNER, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.
- FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-17.
- GELFAND, A.E., et SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- GROENEVELD, R.A., et MEEDEN, G.D. (1977). The mode, median and mean inequality. *American Statistician*, 31, 120-121.
- HARVILLE, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.
- HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.
- HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.
- KASS, R.E., et STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84, 717-726.
- KITANIDIS, P.K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19, 909-921.
- LAIRD, N.M., et LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.
- MARDIA, K.V., et MARSHALL, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.
- MILLER, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics*, 5, 746-762.
- PATTERSON, H.D., et THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.
- PATTERSON, H.D., et THOMPSON, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*. Washington, DC: Biometric Society, 197-207.
- PRASAD, N.G.N., et RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, C.R. (1979). MINQ theory and its relation to ML and MML estimation of variance components. *Sankhyā B*, 41, 138-153.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- ZIMMERMAN, D.L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21, 655-672.
- ZIMMERMAN, D.L., et CRESSIE, N. (1991). Mean-squared prediction error in the spatial linear model. *Annals of the Institute of Statistical Mathematics*, 43, forthcoming.

où $\Delta = \text{diag}\{\delta_1^2, \dots, \delta_n^2\}$. En ajustant le modèle plus général défini par (6.1), (1.5) et (1.6), on pourrait alors vérifier si l'estimation $MVC \sigma_{r_i}^2$ est significativement différente de $\sigma^2 = 1$, ce qui serait un moyen de détecter les erreurs de spécification. (Dans ce cas, il est préférable de recourir à l'estimation MVC plutôt qu'à l'estimation du m.v. puisque tout biais aura une influence considérable sur l'induction pour σ^2 .)

En estimant les paramètres de la matrice de variances par la méthode du maximum de vraisemblance avec contrainte (MVC), on a moins de chances d'obtenir des prédicteurs empiriques de Bayes qui mettent trop de poids sur le modèle de régression (1.5). En contrepartie, on aura une erreur quadratique moyenne de prévision légèrement plus grande. En se servant des propriétés de distribution asymptotique des estimateurs MVC (que l'on vérifie par simulation), on peut aussi obtenir des estimateurs plus précis de l'erreur quadratique moyenne de prévision. En vertu du modèle défini par les expressions (1.4), (1.5) et (1.6), on peut conclure qu'il existe des moyens fiables de faire de l'induction sur les facteurs de redressement $\{F_i : i = 1, \dots, n\}$; les prédicteurs $\{\hat{p}_i(\tilde{Y}; \tilde{\tau}_{r_i}^2) : i = 1, \dots, n\}$ donnent les prédicteurs de l'effectif réel et du sous-dénombrement

$$T_{\text{prd}}^i = \hat{p}_i(\tilde{Y}; \tilde{\tau}_{r_i}^2) C_i \quad \text{et} \quad U_{\text{prd}}^i = 100\{1 - (\hat{p}_i(\tilde{Y}; \tilde{\tau}_{r_i}^2))^{-1}\}; \quad i = 1, \dots, n,$$

respectivement. On peut calculer le biais et l'erreur quadratique moyenne de prévision de ces prédicteurs à l'aide de la méthode (voir Cressie 1991, section 3.2.2).

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à Robert Parker pour l'aide que ce dernier lui a apportée dans sa recherche. Il remercie aussi le rédacteur associé et les deux arbitres qui lui ont fait des commentaires utiles. Cette étude a été rendue possible grâce à la convention sur la statistique no JSA 90-41 intervenue entre le Bureau of the Census des E.-U. et l'université Iowa State. Les conclusions et les opinions formulées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement la position du Bureau of the Census.

BIBLIOGRAPHIE

- CALVIN, J.A., et SEDRANSKY, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-48.
- CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. Dans *Proceedings of Bureau of the Census Fourth Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 127-150.
- CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- CRESSIE, N. (1990). Weighted smoothing of estimated undercount. In *Proceedings of Bureau of the Census 1990 Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 301-325.
- CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.
- CRESSIE, N., et LAHIRI, S.N. (1991). The asymptotic distribution of REML estimators. *Statistical Laboratory Preprint 91-20*, Iowa State University, Ames, Iowa.
- EATON, M.L. (1985). The Gauss-Markov Theorem in multivariate analysis. Dans *Multivariate Analysis - VI*, (Ed. P.R. Krishnaiah). Amsterdam: Elsevier, 177-201.
- ERIKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.

Figure 1c

0	00000001234777799
1	0012334567789
2	0122255688899
3	11223444556889
4	001334445566677788888899
5	00001122233333444445566677788888999
6	000111222222334444444566667777888899999
7	0000001111112222223333444445556677778888899999
8	000000011112223333455555666677778889
9	000111222222333333344444555566688999999
10	00000000111223334444555667778888899
11	00011222333444556667778888899
12	0001111112223334445567789
13	00013333455555566788
14	0001112344445667789
15	00111222344566788
16	0011122223355557999
17	011235566
18	00112566777779
19	117
20	013478
21	123
22	7
23	6
24	
25	02

$$\{\text{var}(t_{2}^{mc})\}^{1/2} \approx 48.73,$$

qui doit être comparé à $S = 45.65$. Enfin, en substituant $\tau^2 = 95.00$ dans (3.29), on obtient

$$\{\text{var}(t_{2}^{pe})\}^{1/2} \approx 50.14,$$

qui doit être comparé à $S = 49.17$.

La simulation nous permet aussi d'étudier les erreurs de prévision "réelles" et d'évaluer le rendement de $M_1(\tau^2)$ et de $M_2(\tau^2)$ *. Si les valeurs des paramètres (5.3) étaient estimées à l'aide des observations initiales, on parlerait alors d'une "bootstrap" paramétrique.

6. CONCLUSIONS ET ANALYSE

La prédiction du sous-dénombrement fondée sur un modèle repose sur une vérification rigoureuse de l'ajustement. Des graphiques diagnostiques basés sur les résidus normalisés ont été proposés à la fin de la section 2 de cet article. Les résidus normalisés $\text{BLUP}, \{Y_i - \hat{p}_i(\tilde{Y}; \hat{\tau}^2)\} / \{[M(\hat{\tau}^2)]^{1/2}\}$, $i = 1, \dots, n$, peuvent eux aussi être utiles. Ils peuvent être utilisés dans un graphique quantile-quantile (voir, par ex. Cressie 1991, p. 255) ou, comme le propose Calvin et Sedransk (1991), être représentés graphiquement en fonction de $\hat{p}_i(\tilde{Y}; \hat{\tau}^2)$; $i = 1, \dots, n$. On pourrait également élargir le modèle (1.4) en y incluant un paramètre de composante de variance inconnu, σ^2 :

$$\tilde{Y} \sim \text{Gau}(\tilde{F}, \sigma^2 \Delta),$$

0	00000037778
1	01111334446679999
2	1114445557778888
3	000022222223333355556666888899999
4	12222222446666777777999999
5	0002222333334455777778888
6	000000011133333444444466667777999999999
7	111122222224444455555577777888888
8	00000222333333555555666666888999
9	1112222244444466667777779999
10	000000022223333355577777788888
11	0000000011111333344444446666777899999
12	11111222222444444444577778
13	000022336666888888999
14	11122244667777999
15	002233558888
16	000001133444777799
17	1222222444555788
18	000033335566899
19	26799
20	02258
21	37
22	11558
23	5
24	79
25	
26	
27	5
28	5
29	2
30	78
31	3
32	
33	2

Figure 1b

Les moyennes (\bar{X}) et les écarts-types des distributions présentées dans la figure 1 sont:

$\hat{\tau}_{ml}^2$	$\bar{X} = 83.56$ $S = 45.65$
$\hat{\tau}_{mm}^2$	$\bar{X} = 96.85$ $S = 57.46$
$\hat{\tau}_{rl}^2$	$\bar{X} = 94.27$ $S = 49.17$

Comparons ces moyennes à la valeur vraie de τ^2 , soit 95.00. Le biais de $\hat{\tau}_{ml}^2$ est évident; $\hat{\tau}_{rl}^2$ a un très faible biais et présente un léger avantage par rapport à $\hat{\tau}_{mm}^2$. Pour ce qui a trait aux écarts-types, $\hat{\tau}_{rl}^2$ est beaucoup plus avantageux que $\hat{\tau}_{mm}^2$ mais désavantageux par rapport à $\hat{\tau}_{ml}^2$. Pour des raisons que nous avons exposées dans la section 3.3, et qui ne sont pas toutes de nature statistique, le biais est plus important que la variance; par conséquent, la méthode du maximum de vraisemblance avec contrainte pourrait remplacer avantageusement la méthode du maximum de vraisemblance pour l'estimation de τ^2 .

On peut vérifier les propriétés de distribution asymptotique de l'estimateur du m. v. et de l'estimateur MVC à l'aide des simulations. (La méthode des moments ne présente aucun intérêt ici car aucune loi de distribution asymptotique n'a encore été définie pour cette méthode.) En substituant $\tau^2 = 95.00$ dans (3.13), on obtient

Cressie (1990) définit des expressions pour déterminer le risque lié à un redressement par $\hat{p}(\bar{Y}; \tau^2)$ et le risque lié à l'absence de redressement. Lorsqu'on substitue $\hat{\tau}^2_{mi}$ et $\hat{g}(\hat{\tau}^2_{mi})$ dans ces expressions, le risque lié à un redressement est de 3,253 tandis que le risque lié à l'absence de redressement est de 34,134. Autrement dit, le fait de ne pas redresser l'effectif recensé accroît le risque de 949% (à la condition que le modèle défini en (1.4), (1.5) et (1.6) soit valable).

5.2 Simulation

Afin de vérifier les propriétés de distribution asymptotique de l'estimateur MVC (et de l'estimateur du m.v.) de τ^2 , nous avons soumis le modèle linéaire décrit dans la section 5.1 à une simulation comprenant les valeurs de paramètres:

(5.3) $\beta_0 = 1.0330, \beta_1 = 0.00712, \beta_5 = -0.000110, \tau^2 = 95.00.$

La simulation

(5.4) $\tilde{Y} \sim \text{Gau}(X\tilde{g}, \Delta + \tau^2 D),$

où Δ est défini en (1.4) (nous avons utilisé les mêmes valeurs $\delta^2_1, \dots, \delta^2_{51}$ que dans la section 5.1 et Cressie (1990)) et D est défini en (1.6), a été exécutée 500 fois et les estimations $\hat{\tau}^2_{mi}, \hat{\tau}^2_{mm}$ et $\hat{\tau}^2_{mi}$ calculées à chaque fois. (Lorsqu'une valeur négative était calculée, elle était ramenée à zéro.) Les diagrammes arborescents des trois séries d'estimations sont présentés dans les figures 1a, 1b et 1c respectivement. Il convient de remarquer que le nombre de zéros est relativement plus élevé dans le cas des estimations du m.v. (figure 1a).

Figure 1: Diagrammes arborescents de la composante de variance estimée τ^2 , selon 500 simulations du modèle (5.4): a) méthode du maximum de vraisemblance (section 3.1); b) méthode des moments (section 3.2) et c) méthode du maximum de vraisemblance avec contrainte (section 3.3).

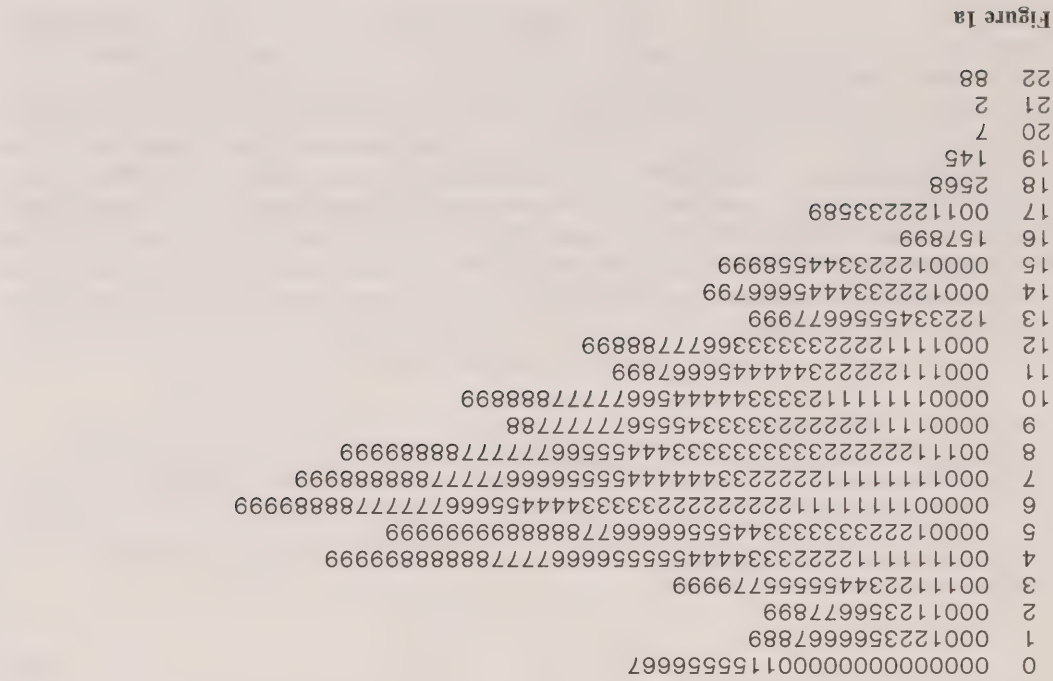


Tableau 1 (fin)

ETAT	Y	m.v.			
		AJUST	POIDS	F12	RMPE1
ala	0.9965	1.0037	0.1190	1.0028	0.00415
aka	1.0288	1.0175	0.4241	1.0223	0.00850
arz	1.0204	1.0157	0.0608	1.0160	0.00448
ark	0.9895	0.9963	0.1161	0.9955	0.00506
cal	1.0307	1.0224	0.0559	1.0228	0.00314
col	1.0033	1.0198	0.1617	1.0171	0.00446
con	0.9886	1.0079	0.0849	1.0063	0.00398
del	0.9938	1.0107	0.4050	1.0039	0.00697
fla	1.0144	1.0120	0.0644	1.0121	0.00271
gga	0.9955	1.0046	0.1368	1.0034	0.00375
hai	1.0111	1.0105	0.2378	1.0106	0.00629
idh	1.0125	1.0070	0.5099	1.0098	0.00507
ill	1.0211	1.0103	0.0967	1.0113	0.00242
ind	0.9936	1.0026	0.1174	1.0015	0.00309
low	0.9932	1.0034	0.1230	1.0021	0.00418
kan	1.0056	1.0091	0.1870	1.0085	0.00432
ky	0.9845	0.9874	0.1264	0.9870	0.00486
lou	1.0234	1.0086	0.0214	1.0089	0.00446
mne	1.0201	0.9993	0.3222	1.0060	0.00557
mld	1.0242	1.0139	0.0583	1.0145	0.00376
mas	0.9882	1.0068	0.1634	1.0037	0.00302
mch	1.0079	1.0081	0.1335	1.0081	0.00242
min	1.0111	1.0049	0.2386	1.0064	0.00339
mis	1.0097	1.0085	0.1060	1.0087	0.00526
mou	1.0080	1.0011	0.1404	1.0021	0.00326
mon	1.0144	1.0059	0.3299	1.0087	0.00656
neb	1.0008	1.0071	0.4587	1.0042	0.00420
nev	1.0265	1.0150	0.2439	1.0178	0.00692
nwh	0.9842	1.0033	0.2646	0.9983	0.00637
nwj	1.0130	1.0105	0.0736	1.0106	0.00283
nwm	1.0236	1.0254	0.2826	1.0249	0.00582
nwy	1.0166	1.0119	0.0663	1.0122	0.00231
noc	1.0118	0.9998	0.0614	1.0005	0.00401
nod	1.0005	0.9970	0.8710	1.0000	0.00310
oho	1.0108	1.0045	0.1055	1.0051	0.00236
okl	0.9977	1.0018	0.1356	1.0013	0.00396
ore	1.0027	1.0088	0.2421	1.0074	0.00408
pen	0.9972	1.0014	0.1227	1.0008	0.00239
rhi	1.0089	0.9940	0.3660	0.9995	0.00591
soc	1.0632	1.0041	0.0176	1.0051	0.00519
sod	1.0008	0.9985	0.7122	1.0002	0.00452
ten	0.9717	0.9967	0.0619	0.9951	0.00413
tex	1.0037	1.0148	0.0393	1.0144	0.00329
uth	1.0040	1.0141	0.3512	1.0105	0.00498
vmt	0.9889	1.0019	0.7901	0.9916	0.00445
vir	1.0009	1.0058	0.1467	1.0051	0.00317
was	1.0142	1.0120	0.1082	1.0123	0.00391
wew	0.9942	0.9879	0.1207	0.9886	0.00567
wis	1.0173	1.0033	0.2461	1.0067	0.00306
wyo	1.0361	1.0127	0.3494	1.0209	0.00829
dcl	1.0375	1.0470	0.1849	1.0452	0.01036
RMPE2					0.01078

Tableau 1

ETAT	Y	AJUST	POIDS	F12	RMPE1	RMPE2
MVC						
ala	0.9965	1.0037	0.1431	1.0026	0.00439	0.00453
aka	1.0288	1.0175	0.4767	1.0229	0.00896	0.00976
arz	1.0204	1.0158	0.0742	1.0162	0.00487	0.00500
ark	0.9895	0.9962	0.1398	0.9953	0.00541	0.00562
cal	1.0307	1.0225	0.0682	1.0231	0.00322	0.00327
col	1.0033	1.0199	0.1926	1.0167	0.00473	0.00495
con	0.9886	1.0079	0.1029	1.0059	0.00435	0.00451
del	0.9938	1.0107	0.4571	1.0030	0.00739	0.00811
fla	1.0144	1.0120	0.0785	1.0122	0.00289	0.00295
gga	0.9955	1.0046	0.1639	1.0031	0.00391	0.00403
hai	1.0111	1.0105	0.2785	1.0107	0.00678	0.00730
idh	1.0125	1.0070	0.5627	1.0101	0.00531	0.00579
ill	1.0211	1.0103	0.1170	1.0116	0.00257	0.00265
ind	0.9936	1.0026	0.1413	1.0013	0.00334	0.00349
low	0.9932	1.0033	0.1478	1.0018	0.00452	0.00475
kan	1.0056	1.0092	0.2215	1.0084	0.00466	0.00496
kyt	0.9845	0.9872	0.1519	0.9868	0.00507	0.00524
lou	1.0234	1.0086	0.0263	1.0090	0.00476	0.00480
mne	1.0201	0.9992	0.3703	1.0093	0.00593	0.00645
mld	1.0242	1.0140	0.0712	1.0147	0.00406	0.00415
mas	0.9882	1.0068	0.1945	1.0032	0.00323	0.00341
mch	1.0079	1.0081	0.1601	1.0081	0.00259	0.00271
min	1.0111	1.0049	0.2793	1.0066	0.00359	0.00383
mis	1.0097	1.0086	0.1279	1.0087	0.00557	0.00575
mon	1.0080	1.0010	0.1681	1.0022	0.00350	0.00367
neb	1.0144	1.0059	0.3785	1.0091	0.00699	0.00761
nev	1.0008	1.0071	0.5117	1.0039	0.00441	0.00480
nwh	1.0265	1.0151	0.2852	1.0183	0.00744	0.00802
nwj	0.9842	1.0033	0.3080	0.9974	0.00684	0.00740
nwm	1.0236	1.0256	0.3276	1.0249	0.00611	0.00648
nwy	1.0166	1.0119	0.0807	1.0123	0.00243	0.00247
noc	1.0118	0.9998	0.0748	1.0007	0.00421	0.00430
nod	1.0005	0.9969	0.8931	1.0001	0.00313	0.00324
oho	1.0108	1.0044	0.1273	1.0052	0.00253	0.00263
okl	0.9977	1.0018	0.1625	1.0011	0.00429	0.00451
ore	1.0027	1.0089	0.2833	1.0071	0.00434	0.00464
pen	0.9972	1.0013	0.1475	1.0007	0.00253	0.00263
rhi	1.0089	0.9939	0.4167	1.0001	0.00625	0.00678
soc	1.0632	1.0040	0.0216	1.0053	0.00555	0.00559
sod	1.0008	0.9985	0.7538	1.0002	0.00464	0.00496
ten	0.9717	0.9966	0.0755	0.9947	0.00439	0.00449
tex	1.0037	1.0149	0.0482	1.0144	0.00341	0.00345
uth	1.0040	1.0142	0.4010	1.0101	0.00524	0.00563
vmt	0.9889	1.0018	0.8232	0.9912	0.00454	0.00479
vir	1.0009	1.0058	0.1753	1.0049	0.00338	0.00354
was	1.0142	1.0121	0.1305	1.0123	0.00418	0.00434
wes	0.9942	0.9877	0.1452	0.9887	0.00603	0.00628
wis	1.0173	1.0032	0.2877	1.0073	0.00325	0.00348
wyo	1.0361	1.0127	0.3992	1.0221	0.00882	0.00963
del	1.0375	1.0474	0.2191	1.0452	0.01081	0.01125

Notons qu'il y a très peu de différence entre les deux séries d'estimations, sauf pour ce qui a trait à τ^2 . Au moyen des estimations du m.v. et des estimations MVC dans $\hat{p}_i(\tilde{X}; \tau^2)$ défini en (2.5), $[M_1(\tau^2)]^{III}$ défini en (2.6), et $[M_2(\tau^2)]^{III}$ défini en (4.7), $i = 1, \dots, n$, nous pouvons calculer les prédicteurs pour petites régions ainsi que la racine carrée des erreurs quadratiques moyennes de prévision estimées. Le tableau 1 donne les résultats pertinents pour les $n = 51$ États; on trouve aussi dans ce tableau le taux de sous-dénombrement brut, Y_i , le résultat du modèle linéaire ajusté, $(X\hat{\beta})_i$, et le poids,

(5.1)

$$w_i \equiv \tau^2 / (C_i \delta_i^2 + \tau^2),$$

de sorte que

(5.2)

$$\hat{p}_i(\tilde{X}; \tau^2) = w_i Y_i + (1 - w_i) (X\hat{\beta})_i, \quad i = 1, \dots, 51.$$

Notons que w_i est systématiquement plus élevé pour l'estimation MVC que pour l'estimation du m.v., ce qui est logique intuitivement compte tenu du biais négatif notablement élevé de τ_{mi}^2 . Ainsi, l'estimation MVC de τ^2 attribue moins de poids au terme de modèle $(X\hat{\beta})_i$, mais d'une manière telle que l'on peut tenir compte de l'effet de l'estimation de τ^2 . Nous notons avec intérêt qu'il y a un inconvéient à utiliser l'estimation MVC; la racine carrée de l'erreur quadratique moyenne de prévision est toujours plus élevée dans ce cas que dans le cas de l'estimation du m.v. Cela n'est pas étonnant si l'on tient compte du fait que l'estimateur MVC est (asymptotiquement) moins efficace que l'estimateur du m.v. Notons en outre que la racine carrée de l'erreur quadratique moyenne de prévision corrigée, $\sqrt{[M_2(\tau^2)]^{III}}$, est de 1 à 9% plus élevée que $\sqrt{[M_1(\tau^2)]^{III}}$.

En ce qui a trait à la prévision, on peut évaluer l'importance globale des deux méthodes d'estimation de τ^2 comparées l'une à l'autre en calculant la somme des carrés pondérée,

$$\sum_{i=1}^{51} \{ \hat{p}_i(\tilde{X}; \tau_{mi}^2) - \hat{p}_i(\tilde{X}; \tau_i^2) \}^2 C_i = 15.$$

Lorsqu'on compare cette somme à,

$$\sum_{i=1}^{51} (Y_i - 1)^2 C_i = 70,421$$

et à

$$\sum_{i=1}^{51} \{ Y_i - \hat{p}_i(\tilde{X}; \tau_{mi}^2) \}^2 = 26,033,$$

on vient bien que, d'un point de vue national, la prévision n'est pas très sensible aux méthodes d'estimation de τ^2 . (Cressie 1990 arrive à la même conclusion après avoir comparé de la même manière les estimateurs τ_{mi}^2 et τ_{mm}^2 .) Cependant, on voit bien aussi d'après le tableau 1 que la racine carrée de l'erreur quadratique moyenne de prévision estimée est beaucoup plus sensible.

Tableau 1: De gauche à droite: les 51 États, identifiés par un code à trois lettres; le taux de sous-dénombrement brut $\{Y_i\}$; le résultat du modèle ajusté $\{(X\hat{\beta})_i\}$; le poids $\{w_i\}$, défini en (5.1); le prédicteur (5.2) (désigné par F12); la racine carrée de l'erreur quadratique moyenne de prévision $\{[M_1(\tau^2)]^{III}\}$ (désignée par RMPE1); et la racine carrée de l'erreur quadratique moyenne de prévision corrigée $\{[M_2(\tau^2)]^{III}\}$ (désignée par RMPE2).

5. COMPARAISON D'ESTIMATEURS AU MOYEN D'UN
EXEMPLE ET D'UNE SIMULATION

5.1 Exemple

Les données de la série PEP 3-8 de l'Enquête postcensitaire de 1980 pour les $n = 51$ "Etats" des E.-U. (y compris Washington, DC) serviront à illustrer la méthode empirique de Bayes. Ces données figurent dans Cressie (1989, tableau 1, colonnes de totaux) et les variances $\delta_1^2, \dots, \delta_{51}^2$ dans (1.3) sont tirées de la colonne de totaux qui a pour titre $MSE_{1/2}$ (et dont les éléments mis au carré sont désignés par MSE_1, \dots, MSE_{51}). En utilisant la relation $F_i = \{1 - U_i/100\}^{-1}$ et la méthode δ_i , nous avons $\delta_i^2 \approx (Y_i^4(MSE_i)/10^4$. Huit variables explicatives, définies par Erickson, Kadane et Tukey (1989), ont été adjointes aux 51 Etats (qui étaient divisés, globalement, en 66 sous-régions comprenant des villes, des portions d'Etat et des Etats entiers). Les variables explicatives étaient les suivantes:

- 1. Pourcentage de minorités.
- 2. Taux de criminalité.
- 3. Pourcentage de la population vivant sous le seuil de la pauvreté.
- 4. Proportion de personnes qui maîtrisent difficilement la langue anglaise.
- 5. Niveau d'instruction.
- 6. Logement.
- 7. Proportion de la population vivant dans l'une ou l'autre de 16 villes centrales définies au préalable.
- 8. Proportion de la population recensée selon la méthode classique.

Afin de trouver parmi ces variables celles qui définiraient un bon modèle du sous-dénombrément, nous nous sommes servis de la méthode de sélection d'Erickson, Kadane et Tukey (1989) mais nous avons pondéré les données proportionnellement à la racine carrée de l'effectif recensé des petites régions. Les variables choisies étaient le pourcentage de minorités, ici, à entrer dans le modèle linéaire, c'est-à-dire que seuls les coefficients de régression β_0, β_1 et β_5 seront ajustés.

Suivant le modèle défini par les expressions (1.4), (1.5) et (1.6), les paramètres inconnus sont $\hat{\beta}$ et τ^2 . D'après l'algorithme de caractérisation (3.8), l'estimation du maximum de vraisemblance de τ^2 est:

$$\tau_{ml}^2 = 47.32,$$

tandis que d'après l'algorithme de caractérisation (3.23), l'estimation MVC de τ^2 est:

$$\tau_{rl}^2 = 58.53.$$

Ces résultats illustrent le phénomène que nous observons plus bas dans la simulation, à savoir que $\tau_{ml}^2 < \tau_{rl}^2$; dans la section 3.3, nous avons donné une explication intuitive de ce phénomène. (Incidentement, Cressie (1990) a calculé la valeur $\tau_{mm}^2 = 94.96$, mais on ne peut dire vraiment s'il existe une relation d'inégalité générale entre les trois estimations.)

Les formules de la section 3 nous ont permis de calculer les estimations suivantes (les erreurs types estimées étant entre parenthèses):

m.v.	MVC
$\hat{\beta}_0 = 1.03227$ (0.00708)	$\hat{\beta}_0 = 1.03246$ (0.00724)
$\hat{\beta}_1 = 0.0006878$ (0.0001402)	$\hat{\beta}_1 = 0.0006941$ (0.0001436)
$\hat{\beta}_5 = -0.001070$ (0.000231)	$\hat{\beta}_5 = -0.001078$ (0.000236)
$\tau^2 = 47.32$ (32.87)	$\tau^2 = 58.53$ (38.1).

Kass et Stefrey (1989) donnent des approximations (de la variance conditionnelle) qui se rapprochent fondamentalement de (4.7), pour des distributions de probabilités qui ne sont pas nécessairement gaussiennes. Cependant, leur approche exige des répétitions indépendantes, ce qui n'est pas une caractéristique des distributions définies en (3.1). Si l'on devait regrouper de petites régions, il est essentiel d'avoir un estimateur approximativement sans biais de tous les éléments de $M_2(\tilde{\gamma})$. On peut facilement généraliser (4.7) par l'expression

$$[M_2(\tilde{\gamma})]_{ij}^* = [M_1(\tilde{\gamma})]_{ij} + 2\text{tr}\{A_{ij}(\tilde{\gamma})B(\tilde{\gamma})\}; i, j = 1, \dots, n,$$

où $A_{ij}(\tilde{\gamma}) \equiv \text{cov}\{\partial p_i(\tilde{\gamma}); \tilde{\gamma}\}/\partial \tilde{\gamma}$, $\partial p_j(\tilde{\gamma})/\partial \tilde{\gamma}$. Prasad et Rao (1990) montrent que, dans le même ordre de grandeur, on peut remplacer $A_{ij}(\tilde{\gamma})$ par $\text{cov}\{\partial p_i^*(\tilde{X})/\partial \tilde{\gamma}, \partial p_j^*(\tilde{X})/\partial \tilde{\gamma}\}$, où $\tilde{p}^*(\tilde{X})$ est défini en (2.1); ces dérivées sont probablement plus faciles à calculer. Lorsque $\tilde{\gamma}$ consiste uniquement en τ^2 dans (1.6), le calcul de $B(\tilde{\gamma})$ est simple; voir (3.13) et (3.26). Considérons maintenant

$$\text{var}(\partial \tilde{p}(\tilde{X}; \tau^2)/\partial \tau^2) = (\partial A(\tau^2)/\partial \tau^2) \Sigma(\tau^2) (\partial A(\tau^2)/\partial \tau^2)', \quad (4.12)$$

où $A(\tau^2)$ est défini en (2.3). Lorsque ce dernier terme est exprimé en fonction de $\Pi(\tau^2)$, défini en (3.27), et de Δ , défini en (1.4), nous avons

$$A(\tau^2) = I - \Delta \Pi(\tau^2). \quad (4.13)$$

Par conséquent, on peut calculer (4.12) à l'aide de (4.13) en se servant des équations (3.4) et (3.5). Alors, $A_{ii}(\tau^2)$, défini en (4.8), est l'élément (i,i) de

$$\Delta(\partial \Pi(\tau^2)/\partial \tau^2) \Sigma(\tau^2) (\partial \Pi(\tau^2)/\partial \tau^2)' \Delta', \quad (4.14)$$

où

$$\begin{aligned} \partial \Pi(\tau^2)/\partial \tau^2 = & -\Sigma(\tau^2) D \Sigma(\tau^2) \{I - X(X'X \Sigma(\tau^2) X'X \Sigma(\tau^2))^{-1} X - \\ & \Sigma(\tau^2) - X(X'X \Sigma(\tau^2) X'X \Sigma(\tau^2))^{-1} D \Sigma(\tau^2) X'X\} \\ & X'X \Sigma(\tau^2) - X(X'X \Sigma(\tau^2) X'X \Sigma(\tau^2))^{-1} X + \Sigma(\tau^2) - X(X'X \Sigma(\tau^2) X'X \Sigma(\tau^2))^{-1} X \end{aligned}$$

(4.15)

il convient de rappeler que $\Sigma(\tau^2) = \Delta + \tau^2 D$ et $D = \text{diag}\{1/C_1, \dots, 1/C_n\}$. On suppose que l'estimateur de l'erreur quadratique moyenne de prévision, $[M_2(\tau^2)]_{ii}^*$, est approximativement non biaisé (Prasad et Rao 1990, avaient étudié un modèle plus particulier que le nôtre). On obtient cet estimateur en combinant les expressions (4.7), (4.14) et (4.10), dans le cas de l'estimation du maximum de vraisemblance, ou (4.7), (4.14) et (4.11), dans le cas de l'estimation MVC. Dans la section qui suit, nous comparerons cet estimateur à l'estimateur courant $[M_1(\tau^2)]_{ii}$ à l'aide de données du recensement et de l'enquête postcensitaire de 1980 aux États-Unis.

Par ailleurs, suivant les hypothèses (4.1) et (4.2) (la caractéristique gaussienne est importante ici, et pourvu que $\tilde{\gamma}$ soit une fonction paire et invariante par translation des observations, on peut se servir des résultats de Harville (1985) pour établir que $M_2(\tilde{\gamma}) - M_1(\tilde{\gamma})$ est définie non négative. (Un estimateur est pair si $\tilde{\gamma}(\tilde{X}) = \tilde{\gamma}(-\tilde{X})$ et est invariant par translation si $\tilde{\gamma}(\tilde{X}) = \tilde{\gamma}(\tilde{X}) + X\tilde{\lambda}$), pour n'importe quel vecteur $p \times 1 \tilde{\lambda}$. Lorsque $\tilde{\gamma}$ consiste uniquement en r^2 dans (1.6), les estimateurs $\hat{\tau}_{mv}^2$, $\hat{\tau}_{mm}^2$ et $\hat{\tau}_{rv}^2$ sont tous pairs et invariants par translation. Intuitivement, l'estimation des paramètres inconnus $\tilde{\gamma}$ engendre des erreurs quadratiques moyennes de prévision plus élevées; les résultats ci-dessus confirment cette intuition.

Cependant, il existe une autre source potentielle de biais du fait que $M_1(\tilde{\gamma})$, et non $M_1(\gamma)$, sert à estimer la matrice de risque. Supposons que l'on choisisse $\tilde{\gamma}$ en vue d'obtenir un estimateur sans biais de la matrice de variances de (\tilde{X}', \tilde{F}') , ce qui, on en conviendra, est un objectif louable. On peut alors se servir des résultats de Eaton (1985) et de Zimmerman et Cressie (1991) pour établir que $M_1(\tilde{\gamma}) - E(M_1(\tilde{\gamma}))$ est définie non négative. (La preuve repose sur une version multidimensionnelle de l'inégalité de Jensen et sur le fait que $\tilde{\rho}(\tilde{X}; \gamma)$, qui peut s'écrire aussi $\Delta(\gamma) \tilde{X}$, minimise la matrice de risque pour tous les prédicteurs linéaires sans biais.)

Etant donné l'expression,

$$M_2(\tilde{\gamma}) - M_1(\tilde{\gamma}) = \{M_2(\tilde{\gamma}) - M_1(\tilde{\gamma})\} + \{M_1(\tilde{\gamma}) - E(M_1(\tilde{\gamma}))\} + \{E(M_1(\tilde{\gamma})) - M_1(\tilde{\gamma})\}, \quad (4.6)$$

les résultats précédents permettent d'établir que la sous-estimation de $M_2(\tilde{\gamma})$ a deux causes. Même si on connaissait une expression pour $M_2(\tilde{\gamma})$, $M_2(\tilde{\gamma})$ serait probablement biaisée pour $M_2(\gamma)$; cela illustre d'ailleurs les difficultés inhérentes à l'estimation des erreurs quadratiques moyennes de prévision.

Prasad et Rao (1990) ont proposé une solution fondée sur un développement asymptotique de $M_2(\tilde{\gamma})$. Envisageons la prédiction du sous-dénombrement dans la région i et désignons par $[M_2(\tilde{\gamma})]^{iii}$ et $[M_1(\tilde{\gamma})]^{iii}$ les éléments (i,i) des matrices de risque $M_2(\tilde{\gamma})$ et $M_1(\tilde{\gamma})$, respectivement. En appliquant rigoureusement la proposition de Prasad et Rao, on obtient l'estimateur de $[M_2(\tilde{\gamma})]^{iii}$,

$$[M_2(\tilde{\gamma})]^{iii} \equiv [M_1(\tilde{\gamma})]^{iii} + 2\text{tr}\{A^{ii}(\tilde{\gamma})B(\tilde{\gamma})\}; i = 1, \dots, n. \quad (4.7)$$

Dans l'équation ci-dessus, $A^{ii}(\tilde{\gamma})$ est une matrice $k \times k$ définie par l'expression

$$A^{ii}(\tilde{\gamma}) = \text{var}\{\partial \rho_i(\tilde{X}; \tilde{\gamma}) / \partial \tilde{\gamma}\} \quad (4.8)$$

et $B(\tilde{\gamma})$ est une matrice qui équivaut précisément ou approximativement à la matrice $k \times k$,

$$E\{(\tilde{\gamma} - \gamma)(\tilde{\gamma} - \gamma)'\}. \quad (4.9)$$

Pour ce qui a trait à l'estimation par la méthode du maximum de vraisemblance,

$$B(\tilde{\gamma}) = J^{-1}, \quad (4.10)$$

où J_γ est défini en (3.7), et en ce qui concerne l'estimation MVC,

$$B(\tilde{\gamma}) = G^{-1}, \quad (4.11)$$

où G_γ est défini en (3.22).

En itérant (3.23) jusqu'à convergence, on obtient l'estimateur MVC $\hat{\gamma}_{rf}$. Cressie et Lahiri (1991) ont démontré que cet estimateur suit approximativement une distribution gaussienne multidimensionnelle ayant comme moyenne $\tilde{\gamma}_{rf}$ et comme matrice de variances asymptotiques

$$G^{-1}_{\gamma} \tag{3.28}$$

Lorsque $\tilde{\gamma}$ consiste uniquement en τ^2 dans (1.6), la matrice (3.28) devient un scalaire,

$$[(\frac{1}{2}) \text{tr} \{ \Pi(\tau^2) D \Pi(\tau^2) D \}]^{-1} . \tag{3.29}$$

Dans la pratique, on calcule les variances et les covariances estimées en évaluant (3.28) à $\tilde{\gamma} = \hat{\gamma}_{rf}$. De plus, l'estimateur par les moindres carrés généralisés (estimés) $\hat{g}(\hat{\gamma}_{rf})$ suit approximativement une distribution gaussienne ayant comme matrice de variances asymptotiques $(X' \Sigma(\tilde{\gamma}) X)^{-1}$.

4. ESTIMATION PLUS PRÉCISE DES ERREURS QUADRATIQUES MOYENNES DE PRÉVISION

Dans cette section, nous allons étudier l'effet que peut avoir sur la prévision l'estimation de $\tilde{\gamma}$ dans $\Sigma(\tilde{\gamma})$, qui figure dans l'expression (3.1). En généralisant (1.5) par

$$\tilde{F} \sim \text{Gau}(X\tilde{g}, \Gamma(\tilde{\gamma})), \tag{4.1}$$

il est clair que

$$\Sigma(\tilde{\gamma}) = \Delta + \Gamma(\tilde{\gamma}). \tag{4.2}$$

En principe, Δ pourrait aussi dépendre de paramètres inconnus (par ex. dans un modèle pour variances d'échantillonnage) et les résultats de cette section seraient tout aussi pertinents. Le prédicteur linéaire non biaisé optimal est

$$\tilde{b}(\tilde{X}; \tilde{\gamma}) = \Gamma(\tilde{\gamma})(\Delta + \Gamma(\tilde{\gamma}))^{-1} \tilde{Y} + \{I - \Gamma(\tilde{\gamma})(\Delta + \Gamma(\tilde{\gamma}))^{-1}\}$$

$$X\{X'(\Delta + \Gamma(\tilde{\gamma}))^{-1}X'(\Delta + \Gamma(\tilde{\gamma}))^{-1}\tilde{X} + \Gamma(\tilde{\gamma})\}^{-1}\tilde{X} \equiv \Delta(\tilde{\gamma})\tilde{X}. \tag{4.3}$$

Alors, la matrice des erreurs quadratiques moyennes de prévision de $b(\tilde{X}; \tilde{\gamma})$, désignée par $M_1(\tilde{\gamma})$, est définie

$$M_1(\tilde{\gamma}) = \Delta(\tilde{\gamma})\Delta\Delta(\tilde{\gamma})' + (\Delta(\tilde{\gamma}) - I)\Gamma(\tilde{\gamma})(\Delta(\tilde{\gamma}) - I)'. \tag{4.4}$$

En réalité, $\tilde{\gamma}$ est inconnu et doit être estimé par $\hat{\gamma}$, par exemple. Le prédicteur empirique de Bayes de \tilde{F} est alors $\tilde{b}(\tilde{X}; \hat{\gamma})$, qui correspond à (4.3) lorsque $\tilde{\gamma} = \hat{\gamma}$. Dans ce cas, $M_1(\hat{\gamma})$ est une mesure inadéquate de la précision du prédicteur; il faudrait utiliser à la place:

$$M_2(\hat{\gamma}) = E\{(\tilde{F} - \tilde{b}(\tilde{X}; \hat{\gamma}))(\tilde{F} - \tilde{b}(\tilde{X}; \hat{\gamma}))'\}. \tag{4.5}$$

C'est cette matrice de risque, ou une estimation de celle-ci, ainsi que le prédicteur $\tilde{b}(\tilde{X}; \hat{\gamma})$ qu'il faudrait connaître. Or, on ne connaît habituellement que $M_1(\hat{\gamma})$; il convient donc de se demander quelles erreurs peuvent découler de l'utilisation de $M_1(\hat{\gamma})$ et s'il n'existerait pas un estimateur de $M_2(\hat{\gamma})$ plus approprié.

L'estimation la plus vraisemblable dépend du profil de la surface de vraisemblance de \tilde{g} et de τ^2 , et à cause de cela, on a plus de chances d'obtenir des valeurs moins élevées pour τ^2 . (Par contraste, on calcule l'estimation MVC en intégrant la fonction de vraisemblance par rapport à \tilde{g} , puis en maximisant le résultat obtenu par rapport à τ^2 . Notons que les adeptes de l'approche bayésienne préconiseraient en plus une intégration par rapport à τ^2 .)

Bien que l'interprétation bayésienne de l'estimation MVC aide à en expliciter les propriétés, $\tilde{\gamma}^{\mu}$ est souvent considéré dans l'optique "fréquentiste" comme un estimateur fondé sur de l'information restreinte.

On peut minimiser (3.19) par rapport à $\tilde{\gamma}$ au moyen de n'importe lequel des algorithmes de gradient. Rappelons-nous que,

$$(3.20) \quad \tilde{W} = A' \tilde{Y}$$

et supposons que A satisfait:

$$AA' = I - X(X'X)^{-1}X' \text{ et } A'A = I.$$

Concentrons-nous pour l'instant sur les ($n - p$) "données" \tilde{W} ; leur distribution conjointe dépend uniquement de $\tilde{\gamma}$ et la fonction de vraisemblance logarithmique négative correspondante (avec contrainte) est $L^W(\tilde{\gamma})$, définie en (3.19).

Définissons le vecteur ($k \times 1$) \tilde{M}_{γ} , qui a pour élément i :

$$(3.21) \quad (\tilde{M}_{\gamma})_i \equiv \partial L^W(\tilde{\gamma}) / \partial \gamma_i = (\frac{1}{2} \text{tr} \{ \Pi(\tilde{\gamma}) \Sigma_i(\tilde{\gamma}) \} - (\frac{1}{2} \tilde{Y}' \Pi(\tilde{\gamma}) \Sigma_i(\tilde{\gamma}) \tilde{Y}),$$

et la matrice ($k \times k$) G_{γ} , qui a pour élément (i, j):

$$(3.22) \quad (G_{\gamma})_{ij} \equiv E(\partial^2 L^W(\tilde{\gamma}) / \partial \gamma_i \partial \gamma_j) = (\frac{1}{2} \text{tr} \{ \Pi(\tilde{\gamma}) \Sigma_i(\tilde{\gamma}) \Pi(\tilde{\gamma}) \Sigma_j(\tilde{\gamma}) \},$$

où $\Pi(\tilde{\gamma})$ est défini au-dessous de l'équation (3.19) et $\Sigma_i(\tilde{\gamma})$ est défini en (3.5). (Les expressions (3.21) et (3.22) sont attribuables à Harville (1977).) Alors, l'algorithme de Gauss-Newton (caractérisation) servant à déterminer $\tilde{\gamma}^{\mu}$ est:

$$(3.23) \quad \tilde{\gamma}_{(t+1)} = \tilde{\gamma}_{(t)} - (G_{(t)}^{\gamma})^{-1} \tilde{M}_{(t)}^{\gamma},$$

où $G_{(t)}^{\gamma}$ et $\tilde{M}_{(t)}^{\gamma}$ désignent respectivement G_{γ} et \tilde{M}_{γ} , évaluées à $\tilde{\gamma}_{(t)}$.

Lorsque $\tilde{\gamma}$ consiste uniquement en τ^2 dans (1.6), l'algorithme (3.23) est particulièrement simple. Dans la simulation et l'exemple présentés dans la section 5, nous avons utilisé la valeur de départ (3.9). Alors, (3.23) est,

$$(3.24) \quad (\tau^2)_{(t+1)} = (\tau^2)_{(t)} - (G_{(t)}^{\tau})^{-1} M_{(t)}^{\tau},$$

où

$$(3.25) \quad M_{\tau}^{\tau} = (\frac{1}{2} \text{tr} \{ \Pi(\tau^2) D \} - (\frac{1}{2} \tilde{Y}' \Pi(\tau^2) D \Pi(\tau^2) \tilde{Y}),$$

$$(3.26) \quad G_{\tau}^{\tau} = (\frac{1}{2} \text{tr} \{ \Pi(\tau^2) D \Pi(\tau^2) D \}),$$

$$(3.27) \quad \Pi(\tau^2) = \Sigma(\tau^2)^{-1} - \Sigma(\tau^2)^{-1} X (X' \Sigma(\tau^2)^{-1} X)^{-1} X' \Sigma(\tau^2)^{-1},$$

sont évaluées à $\tau^2 = (\tau^2)_{(t)}$. Rappelons-nous, en outre, que $\Sigma(\tau^2) = \Delta + \tau^2 D$ et $D = \text{diag} \{ 1/C_1, \dots, 1/C_n \}$.

MVM – maximum de vraisemblance marginal – en ce qui regarde l'estimation des composantes de variance. Récemment, des auteurs l'ont appelée méthode du maximum de vraisemblance résiduel, bien qu'ils aient conservé le sigle REML en anglais (restricted maximum likelihood – si $E(\tilde{g}'\tilde{X}) = 0$ pour tous \tilde{g} et \tilde{X} ; ainsi, $\tilde{g}'\tilde{X}$ est un contraste d'erreur si et seulement si $\tilde{g}'X = \tilde{0}$). Soit $\tilde{W} = A'\tilde{X}$ un vecteur de $n - p$ contrastes d'erreur linéairement indépendants, c'est-à-dire que les $(n - p)$ colonnes de A sont linéairement indépendantes et $A'X = 0$. Suivant l'hypothèse gaussienne (3.1), $\tilde{W} \sim \text{Gau}(\tilde{0}, A'\Sigma(\tilde{Y})A)$, qui ne dépend pas de \tilde{g} . La fonction de vraisemblance logarithmique négative est donc,

$$L^W(\tilde{Y}) = ((n - p)/2)\log(2\pi) + (1/2)\log(|A'\Sigma(\tilde{Y})A|) + (1/2)\tilde{W}'(A'\Sigma(\tilde{Y})A)^{-1}\tilde{W}.$$

Si l'on définit \tilde{W} , au moyen d'une autre série de $(n - p)$ contrastes linéairement indépendants, la nouvelle fonction de vraisemblance logarithmique négative ne différerait de $\tilde{L}^W(\tilde{Y})$ que par une constante additive (Harville 1974). En effet, pour la matrice A qui satisfait l'équation $AA' = I - X(X'X)^{-1}X'$ (et $A'A = I$),

$$L^W(\tilde{Y}) = ((n - p)/2)\log(2\pi) - (1/2)\log(|X'X|) + (1/2)\log(|\Sigma(\tilde{Y})|) + (1/2)\log(|X'\Sigma(\tilde{Y})X|) + (1/2)\tilde{Y}'\Pi(\tilde{Y})\tilde{Y}, \quad (3.19)$$

où $\Pi(\tilde{Y}) = \Sigma(\tilde{Y})^{-1} - \Sigma(\tilde{Y})^{-1}X(X'X)^{-1}X'\Sigma(\tilde{Y})^{-1}$, voir Harville (1974). On obtient une estimation du maximum de vraisemblance avec contrainte de \tilde{Y} , désignée par \tilde{Y}^r , en minimisant (3.19) par rapport à \tilde{Y} . La distinction entre l'estimation par la méthode MVC et l'estimation par rapport à n .

La méthode MVC a été conçue dans le but d'estimer les paramètres des composantes de la variance; c'est dans cette perspective qu'ont été élaborés des algorithmes numériques (Harville 1977), des ajustements robustes (Fellner 1986) et des lois de distribution (Cressie et Lahiri 1991). Kitaniadis (1983) et Zimmermann (1989) donnent les détails mathématiques de la minimisation de (3.19) par itération.

Harville (1974) donne une justification bayésienne de la méthode MVC en supposant une distribution *a priori* non informative pour \tilde{g} , qui est statistiquement indépendant de \tilde{Y} , et en montrant que la distribution *a posteriori* marginale de \tilde{Y} est proportionnelle au produit de (3.19) par la distribution *a priori* de \tilde{Y} . Lorsque cette dernière est non informative, les estimations MVC correspondent aux estimations *a posteriori* maximum marginales. Par conséquent, lorsque les distributions *a priori* non informatives de \tilde{g} et de \tilde{Y} sont indépendantes, l'estimation par la méthode MVC peut être vue comme un compromis entre l'estimation par la méthode du maximum de vraisemblance et l'estimation bayésienne avec fonction quadratique de perte. En ce qui a trait au modèle défini par (1.4), (1.5) et (1.6), la seconde méthode donnerait une estimation de Bayes, $\int_0^\infty \tau^2 \exp\{-L^W(\tau^2)\}d\tau^2$, que l'on peut aussi obtenir en **moyennant sur** τ^2 , pondéré par la fonction de vraisemblance **complète**, $\exp\{-L(\tilde{g}, \tau^2)\}$. Par ailleurs, la méthode du maximum de vraisemblance produit comme estimation de τ^2 la valeur τ_{mc}^2 , que l'on obtient en **maximisant** la fonction de vraisemblance **complète**. Quant à la méthode MVC, elle fait la moyenne de la fonction de vraisemblance complète par rapport à \tilde{g} mais maximise la fonction de vraisemblance résultante (avec contrainte) par rapport à τ^2 . L'estimation du maximum de vraisemblance de τ^2 tend à être biaisée vers zéro parce que la fonction de vraisemblance, en tant que fonction de τ^2 , est étalée vers la droite. Lorsque cette fonction est normalisée de manière à ce que son intégrale soit un, la moyenne de cette fonction est généralement plus grande que le mode (voir, par ex., Groeneveld et Meeden 1977).

où $\Pi^U \equiv U^{-1} - U^{-1}X(X'U^{-1}X)^{-1}X'U^{-1}$. En supposant que $\Sigma(\tilde{\gamma}) = \Delta + \gamma_1 \Gamma_1 + \dots + \gamma_k \Gamma_k$, où les Γ_i sont connus, on obtient,

$$\sum_k \gamma_i \text{tr}(\Gamma_i \Pi^U) = E(\tilde{e}^U \tilde{e}^U) - \text{tr}(\Delta \Pi^U).$$

En choisissant k U_j différents, $j = 1, \dots, k$ (par ex.: U_1, U_2, \dots, U_k^1), on obtient k équations avec k inconnues:

$$(3.16) \quad \sum_k \gamma_i \text{tr}(\Gamma_i \Pi^{U_j}) = \tilde{e}_{U_j}' \tilde{e}_{U_j} - \text{tr}(\Delta \Pi^{U_j}); j = 1, \dots, k,$$

lesquelles peuvent être résolues en $\hat{\gamma}_1, \dots, \hat{\gamma}_k$. Il est essentiel de vérifier si la solution $\hat{\gamma}$ appartient à l'espace des paramètres $\{\tilde{\gamma}: \sum_{i=1}^k \gamma_i \Gamma_i \text{ est définie positive}\}$.

Lorsque $\tilde{\gamma}$ consiste uniquement en τ^2 dans (1.6), une seule matrice U dans (3.16) est nécessaire. Avec les prédicteurs antérieurs du sous-dénominateur, l'estimation de τ^2 reposait sur $U = I$ (Erickson et Kadane 1985; Freedman et Navidi 1986; Erickson, Kadane et Tukey 1989); cependant, une petite étude de sensibilité concernant le modèle hétéroscédastique (1.6) a permis de croire qu'il existait un meilleur estimateur.

Choisissons $U_\alpha = \Delta + \Gamma(\alpha)$ dans (3.15) comme une reproduction du modèle (1.7). Alors, quand $\alpha = \tau^2$ (la valeur vraie), Fay et Herriot (1979) montrent que

$$(3.17) \quad E(\tilde{e}_{U_\alpha}' \tilde{e}_{U_\alpha}) = n - p,$$

où n est le nombre de régions, p est le nombre de variables prédictives dans la matrice X (par ex.: $p = 3$ pour le modèle utilisé dans la section 5) et \tilde{e}_{U_α} est le résidu normalisé défini en (3.14). Par conséquent, l'estimateur par la méthode des moments proposé pour τ^2 est la valeur de

$$(3.18) \quad \tilde{e}_{U_\alpha}' \tilde{e}_{U_\alpha} = n - p,$$

cette équation peut être résolue à l'aide de la méthode itérative de Newton-Raphson ou d'une méthode de fractionnement simple; désignons l'estimateur en question par $\hat{\tau}_{mm}^2$.

Fay et Herriot (1979) notent que les estimateurs $\hat{\tau}_{mm}^2$ et $\hat{\tau}_{mm}^2$ se distinguent l'un de l'autre surtout par la façon dont les régions avec de faibles valeurs δ_j^2 sont pondérées dans le processus d'estimation; pour de telles régions, les carrés des résidus ont relativement plus de poids avec $\hat{\tau}_{mm}^2$ qu'avec $\hat{\tau}_{mm}^2$. Compte tenu de cette observation et des résultats d'une petite étude de simulation sur le biais, Cressie (1990) a donné la préférence à $\hat{\tau}_{mm}^2$. Néanmoins, du point de vue asymptotique, $\hat{\tau}_{mm}^2$ est parfaitement efficace et répond à une loi de distribution connue. En revanche, le fait qu'il n'existe aucune loi de distribution (asymptotique) particulière pour $\hat{\tau}_{mm}^2$ pose des problèmes propres à ce paramètre, par ex.: comment faire de l'inférence sur $\hat{\tau}_{mm}^2$ et comment corriger l'erreur quadratique moyenne de prévision (section 4). Nous proposons ci-dessous un estimateur plus acceptable, supérieur à l'estimateur du maximum de vraisemblance au point du vue du biais.

3.3 Estimation par la méthode du maximum de vraisemblance avec contrainte (MVC)

Il s'agit, ici, de trouver un estimateur convenable des paramètres de la matrice de variances $\tilde{\gamma}$, de (3.1). La méthode du maximum de vraisemblance avec contrainte (MVC), élaborée par Patterson et Thompson (1971, 1974), applique le principe du maximum de vraisemblance aux

Alors (3.8) est,

$$(\tau_2)_{(\ell+1)} = (\tau_2)_{(\ell)} - \{ (\tfrac{1}{2}) \} \sum_n^{l=1} 1/(C_l \delta_2^l + (\tau_2)_{(\ell)}^2) \{ L_{-1}^{\tau_2} \}_2 \}; \ell = 0, 1, \dots \quad (3.10)$$

où

$$L_{(\ell)}^{\tau_2} = (\tfrac{1}{2}) \sum_n^{l=1} 1/(C_l \delta_2^l + (\tau_2)_{(\ell)}^2)$$

$$- (\tfrac{1}{2}) \{ \bar{Y} - X \hat{g}((\tau_2)_{(\ell)}) \}' \text{diag} \{ C_l / (C_l \delta_2^l + (\tau_2)_{(\ell)}^2) \} \{ \bar{Y} - X \hat{g}((\tau_2)_{(\ell)}) \} \}. \quad (3.11)$$

En itérant (3.8) jusqu'à convergence, on obtient l'estimateur du maximum de vraisemblance $\hat{\gamma}_{mt}$, qui, une fois substitué dans l'équation (3.3), donne l'estimateur du maximum de vraisemblance $\hat{g}(\hat{\gamma}_{mt})$. Suivant des conditions de régularité appropriées (voir, par ex. Mardia et Marshall 1984), $(\hat{g}(\hat{\gamma}_{mt})', \hat{\gamma}_{mt}')'$ suit approximativement une distribution gaussienne multidimensionnelle qui a pour moyenne $(\tilde{g}', \tilde{\gamma}')'$ et pour matrice de variances asymptotiques

$$\begin{bmatrix} (X' \Sigma(\tilde{\gamma})^{-1} X)^{-1} & 0 \\ 0 & J_{-1}^{\tilde{\gamma}} \end{bmatrix}; \quad (3.12)$$

lorsque $\tilde{\gamma}$ consiste uniquement en τ_2 dans (1.6), la matrice (3.12) devient,

$$\begin{bmatrix} 0 & \left\{ (\tfrac{1}{2}) \sum_n^{l=1} 1/(C_l \delta_2^l + \tau_2^2) \right\}^{-1} \\ (X' \Sigma(\tau_2)^{-1} X)^{-1} & 0 \end{bmatrix}. \quad (3.13)$$

En pratique, on obtient les variances et les covariances estimées en évaluant (3.12) en fonction de l'estimation la plus vraisemblable $\tilde{\gamma}_{mt}$.

3.2 Estimation par la méthode des moments

Il n'existe pas d'estimateur par la méthode des moments particulier pour $\tilde{\gamma}$, l'essentiel est d'apparier les moments d'ordre inférieur des observations avec les moments empiriques correspondants. Si on n'utilise que les moments du premier ordre et du deuxième ordre, il est clair que l'hypothèse gaussienne en (3.1) est superflue. Soit U une matrice symétrique définie positive. Considérons l'estimateur par régression pondéré, $\hat{g}_U \equiv (X' U^{-1} X)^{-1} X' U^{-1} \bar{Y}$, et les résidus pondérés,

$$\tilde{e}_U \equiv U^{-1/2} (I - X(X' U^{-1} X)^{-1} X' U^{-1}) \bar{Y}. \quad (3.14)$$

Alors, par des opérations simples de l'algèbre matricielle, nous avons

$$E(\tilde{e}_U \tilde{e}_U') = \text{tr}(\Sigma(\tilde{\gamma}) \Pi_U), \quad (3.15)$$

(3.2)
$$\Sigma(\tilde{\gamma}) = \Delta + \Gamma(\tau^2),$$
 où $\tilde{\gamma}$ consiste en un seul paramètre, τ^2 .

Si $\tilde{\gamma}$ est connu, on peut estimer facilement \tilde{g} :

(3.3)
$$\tilde{g}(\tilde{\gamma}) \equiv (X' \Sigma(\tilde{\gamma})^{-1} X)^{-1} X' \Sigma(\tilde{\gamma})^{-1} \tilde{Y}.$$

Dans la réalité toutefois, $\tilde{\gamma}$ est inconnu et il faut l'estimer; en substituant son estimateur dans (3.3), on obtient un estimateur par les moindres carrés généralisés estimés de \tilde{g} . Dans le reste de cette section, nous examinons trois méthodes d'estimation de $\tilde{\gamma}$.

3.1 Estimation par la méthode du maximum de vraisemblance

La fonction de vraisemblance logarithmique négative de \tilde{g} et $\tilde{\gamma}$ est:

(3.4)
$$L(\tilde{g}, \tilde{\gamma}) = (n/2) \log(2\pi) + (1/2) \log(|\Sigma(\tilde{\gamma})|) + (1/2) (\tilde{Y} - X\tilde{g})' \Sigma(\tilde{\gamma})^{-1} (\tilde{Y} - X\tilde{g}).$$

En minimisant cette fonction, on obtient les estimations les plus vraisemblables \hat{g}_{ml} et $\hat{\gamma}_{ml}$. La partie délicate de cette minimisation est de déterminer $\hat{\gamma}_{ml}$. L'algorithme de Gauss-Newton (ou de caractérisation) est décrit notamment dans Harville (1977) et dans Mardia et Marshall (1984), et nous le reproduisons ici dans tous ses détails.

Définissons,

(3.5)
$$\begin{aligned} \Sigma_i(\tilde{\gamma}) &\equiv \partial \Sigma(\tilde{\gamma}) / \partial \gamma_i; i = 1, \dots, k, \\ \Sigma'_i(\tilde{\gamma}) &\equiv \partial \Sigma^{-1}(\tilde{\gamma}) / \partial \gamma_i = - \Sigma(\tilde{\gamma})^{-1} \Sigma'_i(\tilde{\gamma}) \Sigma(\tilde{\gamma})^{-1}; i = 1, \dots, k, \end{aligned}$$

le vecteur \tilde{L}_{γ} , de dimension $k \times 1$, dont l'élément i est défini:

(3.6)
$$(\tilde{L}_{\gamma})_i \equiv (1/2) \text{tr}(\Sigma(\tilde{\gamma})^{-1} \Sigma'_i(\tilde{\gamma})) + (1/2) (\tilde{Y} - X\tilde{g})' \Sigma'_i(\tilde{\gamma}) (\tilde{Y} - X\tilde{g}),$$

et la matrice J_{γ} , de dimension $k \times k$, dont l'élément (i,j) est défini:

(3.7)
$$(J_{\gamma})_{ij} \equiv (1/2) \text{tr}(\Sigma(\tilde{\gamma})^{-1} \Sigma'_i(\tilde{\gamma}) \Sigma(\tilde{\gamma})^{-1} \Sigma'_j(\tilde{\gamma})).$$

Alors,

(3.8)
$$\tilde{\gamma}_{(t+1)} = \tilde{\gamma}_{(t)} - (J_{(t)}^{\gamma})^{-1} \tilde{L}_{(t)}^{\gamma},$$

où $J_{(t)}^{\gamma}$ et $\tilde{L}_{(t)}^{\gamma}$ désignent respectivement J_{γ} et \tilde{L}_{γ} , évalués à $\tilde{\gamma} = \tilde{\gamma}_{(t)}$ et à $\tilde{g} = \tilde{g}_{(t)}$.

Lorsque $\tilde{\gamma}$ consiste uniquement en τ^2 dans (1.6), l'algorithme (3.8) est particulièrement simple. Dans la simulation et l'exemple présentés dans la section 5, nous avons utilisé la valeur de départ

(3.9)
$$\begin{aligned} (\tau^2)_{(0)} &\equiv \{1/(n - p)\} (\tilde{Y} - X(X'D^{-1}X)^{-1}X'D^{-1}\tilde{Y})' D^{-1} \\ &\quad (\tilde{Y} - X(X'D^{-1}X)^{-1}X'D^{-1}\tilde{Y}). \end{aligned}$$

dans (2.1), ce qui donne

$$\tilde{b}(\tilde{Y}; \tau^2) = \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\tilde{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}X\{X'(\Delta + \Gamma(\tau^2))^{-1}X' - I\}^{-1}\tilde{Y} \equiv \Lambda(\tau^2)\tilde{Y} \quad (2.3)$$

(Cressie 1990). La matrice des erreurs quadratiques moyennes de prévision est,

$$M_1(\tau^2) \equiv E\{(\tilde{F} - \tilde{b}(\tilde{Y}; \tau^2))(\tilde{F} - \tilde{b}(\tilde{Y}; \tau^2))'\} = \Lambda(\tau^2)\Delta\Lambda(\tau^2)' + (\Lambda(\tau^2) - I)\Gamma(\tau^2)(\Lambda(\tau^2) - I)' \quad (2.4)$$

De manière plus réaliste, τ^2 est aussi inconnu. On obtient un **prédicteur empirique de Bayes** en substituant un estimateur $\hat{\tau}^2$ dans $\Lambda(\tau^2)$, ce qui donne

$$\tilde{b}(\tilde{Y}; \hat{\tau}^2) = \Lambda(\hat{\tau}^2)\tilde{Y}. \quad (2.5)$$

On voit facilement que lorsque $\hat{\tau}^2$ est l'estimateur du maximum de vraisemblance de τ^2 , l'équation (2.5) est l'estimateur du maximum de vraisemblance du prédicteur de Bayes.

Le prédicteur (2.5) a été proposé par Ericsen et Kadane (1985), puis critiqué par Freedman et Navidi (1986). À propos, les prédicteurs de Freedman et Navidi peuvent sembler différents des prédicteurs (2.1), (2.3) et (2.5) mais en réalité, ils leur sont identiques si l'on tient compte de l'équation $A(A + B)^{-1}B = (A^{-1} + B^{-1})^{-1}$, où A et B sont des matrices carrées telles que A , B et $A + B$ ont des inverses.

En substituant $\hat{\tau}^2$ dans (2.4), on obtient un estimateur de la matrice des erreurs quadratiques moyennes de prévision:

$$M_1(\hat{\tau}^2) \equiv \Lambda(\hat{\tau}^2)\Delta\Lambda(\hat{\tau}^2)' + (\Lambda(\hat{\tau}^2) - I)\Gamma(\hat{\tau}^2)(\Lambda(\hat{\tau}^2) - I)' \quad (2.6)$$

Comme (2.6) ne tient pas compte de l'estimation de τ^2 dans $\hat{b}(\tilde{Y}; \hat{\tau}^2)$, il risque d'être un estimateur biaisé de $E\{(\tilde{F} - \tilde{b}(\tilde{Y}; \hat{\tau}^2))(\tilde{F} - \tilde{b}(\tilde{Y}; \hat{\tau}^2))'\}$. Nous traitons plus en détail cette importante question dans la section 4.

Une fois calculés \hat{g} et $\hat{\tau}^2$, on peut faire un diagnostic de modèle pour vérifier l'ajustement du modèle estimé. Par exemple, un graphique quantile-quantile mettant en relation les résidus normalisés $(\Delta + \Gamma(\hat{\tau}^2))^{-1/2}(\tilde{Y} - X\hat{g})$ et les statistiques d'ordre théoriques d'une distribution gaussienne d'unités n'a révélé aucun manque d'ajustement apparent pour le modèle utilisé dans la section 5. Le diagnostic de modèle est traité plus en détail dans la section 6.

3. ESTIMATION DES PARAMÈTRES DE LA MATRICE DE VARIANCES

Dans cette section, nous supposons le modèle linéaire général,

$$\tilde{Y} \sim \text{Gau}(X\tilde{g}, \Sigma(\tilde{Y})), \quad (3.1)$$

où \tilde{Y} est le vecteur $k \times 1$ des paramètres de la matrice de variances. En particulier, le modèle défini par (1.4), (1.5) et (1.6) donne,

variance Δ et $\Gamma(\tau^2)$ ne renferment qu'un seul paramètre inconnu, soit τ^2 . Il est utile de souligner ici que les méthodes élaborées dans cet article peuvent être facilement généralisées au delà de ce simple problème de composantes de la variance. Le modèle linéaire général est considéré dans la section 3.

Dans la section 2, nous définissons le prédicteur de Bayes et le prédicteur empirique de Bayes de \tilde{F} . L'estimation de \tilde{g} est simple mais il y a plusieurs manières d'estimer τ^2 . Dans la section 3, nous présentons trois méthodes d'estimation particulières: la méthode du maximum de vraisemblance (m.v.), la méthode des moments et la méthode du maximum de vraisemblance avec contrainte (MVC). Dans la section suivante, nous analysons l'effet de l'estimation de τ^2 sur l'erreur quadratique moyenne de prévision. Enfin, nous comparons les diverses méthodes par une simulation et un exemple dans la section 5 et présentons nos conclusions et une analyse dans la section 6.

2. PRÉDICTION EMPIRIQUE DE BAYES

Dans cet article, l'effectif réel de la population d'une petite région est réputé inconnu. Une fois que l'effectif recensé est connu, on révise le degré d'incertitude qui avait été associé à l'origine au "chiffre réel de population". Les modèles statistiques du sous-dénombrement **dépendent** donc des chiffres du recensement. Le modèle que nous avons défini dans la section 1 au moyen des expressions (1.4), (1.5) et (1.6) servira dans les sections 2, 3 et 4. En utilisant un équivalent matriciel de la fonction quadratique de perte, on a pour prédicteur optimal $E(\tilde{F} | \tilde{Y})$ (Cressie 1990), qui est,

$$\tilde{p}^*(\tilde{Y}) \equiv \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\tilde{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}X\tilde{g} \tag{2.1}$$

et la matrice des erreurs quadratiques moyennes de prévision est,

$$E\{(\tilde{F} - \tilde{p}^*(\tilde{Y}))(\tilde{F} - \tilde{p}^*(\tilde{Y}))'\} = \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}\Gamma(\tau^2). \tag{2.2}$$

Pour ce qui a trait à la matrice de perte, $L(\tilde{F}, \tilde{p}) \equiv (\tilde{F} - \tilde{p})(\tilde{F} - \tilde{p})'$, on voit facilement que (2.1) est un prédicteur de **Bayes** de \tilde{F} . De fait, \tilde{g} et τ^2 sont inconnus et l'expression (2.1) n'est donc pas une statistique (c'est-à-dire qu'elle n'est pas strictement une fonction des observations). Selon l'approche bayésienne, il serait indiqué ici de définir d'autres distributions et hyper-distributions *a priori* pour tous les paramètres inconnus. (Cette solution au problème des paramètres inconnus est parfois appelée méthode hiérarchique de Bayes et exige une connaissance préalable de la variabilité de processus que de nombreux scientifiques estiment ne pas avoir. Néanmoins, on peut souvent obtenir des estimateurs raisonnables avec les distributions et hyper-distributions *a priori* non informatives.) Il arrive souvent que les distributions *a posteriori* posent des difficultés insolubles sur le plan analytique. Si le modèle et la distribution *a priori* étaient définis en fonction de leurs distributions conditionnelles, on pourrait se servir du modèle de Gibbs pour obtenir numériquement toutes les distributions marginales et les distributions conjointes voulues (voir, par exemple, Gelfand et Smith 1990).

La solution proposée dans cet article consiste à tenir tous les paramètres, sauf \tilde{F} , pour fixes mais inconnus et à utiliser les observations \tilde{Y} pour estimer ces paramètres. C'est ce qu'on appelle la méthode **empirique de Bayes**. Bien que nous supposons dans cet article une distribution *a priori* paramétrique (conjuguée), nous pourrions aussi bien utiliser une distribution *a priori* non paramétrique (voir, par exemple, Laird et Louis 1987). Supposons maintenant que \tilde{g} est inconnu mais que τ^2 est (pour l'instant) connu. En utilisant encore une fois l'équivalent matriciel de la fonction quadratique de perte, on obtient le prédicteur linéaire non biaisé optimal en substituant l'estimateur par les moindres carrés généralisés:

et Kadane 1985, Freedman et Navidi 1986 et Cressie 1988.) Des États comme la Californie, le Texas et New York profiteraient largement d'un redressement en fonction du sous-énumbrement, c.-à-d. de la substitution de $F_i C_i$ à C_i où F_i est un *facteur de redressement*. Le facteur de redressement approprié est

$$(1.2) \qquad F_i = T_i/C_i,$$

et ce facteur est lié au sous-dénombrement par la relation suivante,

$$F_i = \{1 - U_i/100\}^{-1}.$$

L'équation (1.2) se prête mal à un redressement car on ne connaît pas l'effectif réel de la population, T_i . Afin d'obtenir de l'information supplémentaire qui permettra d'estimer F_i , le Census Bureau des E.-U. effectue une enquête post-censitaire (EP) qui permet de savoir si certaines personnes, en l'occurrence celles incluses dans l'échantillon de l'EP, ont été recensées ou non (voir, par exemple, Wolter 1986). L'enquête implique la participation de plusieurs centaines de milliers de ménages et produit des facteurs de redressement "bruts" $\{Y_i : i = 1, \dots, n\}$ qui doivent être lissés.

Supposons que, étant donné F_i ,

$$(1.3) \qquad Y_i \sim \text{Gau}(F_i, \delta_i^2),$$

c'est-à-dire que Y_i , moyennant F_i , suit une distribution gaussienne de moyenne F_i et de variance δ_i^2 . Si on pose en plus l'hypothèse d'indépendance, on obtient

$$(1.4) \qquad \tilde{Y} \sim \text{Gau}(\tilde{F}, \Delta),$$

où $\tilde{Y} \equiv (Y_1, \dots, Y_n)'$, $\tilde{F} \equiv (F_1, \dots, F_n)'$ et Δ est la matrice diagonale $n \times n$ diag $\{\delta_1^2, \dots, \delta_n^2\}$.

Supposons maintenant que

$$(1.5) \qquad \tilde{F} \sim \text{Gau}(X\tilde{\theta}, \Gamma(\tau^2)),$$

où X est une matrice $n \times p$ de variables explicatives, $\tilde{\theta}$ est un vecteur $p \times 1$ de coefficients (inconnus) du modèle linéaire, $\Gamma(\tau^2)$ est une matrice diagonale $n \times n$:

$$(1.6) \qquad \Gamma(\tau^2) \equiv \tau^2 D$$

et $D \equiv \text{diag}\{1/C_1, \dots, 1/C_n\}$. Le modèle hétéroscédastique défini par les expressions (1.5) et (1.6) est analysé en profondeur dans Cressie (1990). Intuitivement, il est raisonnable de penser que dans le cas des régions à forte population, le facteur de redressement a une variance moins élevée; Cressie (1989) en fait la démonstration tant d'un point de vue bayésien que "fréquentiste".

Le modèle défini par les expressions (1.4) et (1.5) peut aussi s'écrire:

$$(1.7) \qquad \tilde{Y} = X\tilde{\theta} + \tilde{v} + \varepsilon,$$

où les vecteurs $n \times 1$ \tilde{v} et ε sont statistiquement indépendants, $\tilde{v} \sim \text{Gau}(\tilde{0}, \Gamma(\tau^2))$ et $\varepsilon \sim \text{Gau}(\tilde{0}, \Delta)$. Si nous supposons maintenant que les valeurs $\delta_1^2, \dots, \delta_n^2$ sont calculées à l'aide de formules de la variance d'échantillonnage adaptées à la base de sondage de l'EP, il ne reste plus que les paramètres $\tilde{\theta}$ et τ^2 à estimer. Par conséquent, les deux composantes de

Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes

NOEL CRESSIE¹

RÉSUMÉ

Une façon de calculer le sous-dénombrement au niveau infra-national (par ex.: pour un État) est de prendre des données-échantillon d'une enquête postcensitaire et de les lisser suivant un modèle linéaire de variables explicatives. Le rapport entre la variance de l'erreur d'échantillonnage et la variance de l'erreur de modèle correspondante détermine le degré de lissage. L'estimation par la méthode du maximum de vraisemblance peut mener à un lissage excessif et, par conséquent, rendre le calcul du sous-dénombrement trop tributaire du modèle linéaire. Les estimateurs du maximum de vraisemblance avec contrainte (MVC) ne présentent pas de tels inconvénients. Dans cet article, on traite la prévision empirique de Bayes du sous-dénombrement fondée sur l'estimation MVC et on la compare, par des exemples et des simulations, à celle fondée sur la méthode du maximum de vraisemblance et à celle fondée sur une méthode des moments. Les propriétés de distribution pour grand échantillon des estimateurs MVC permettent un calcul précis de l'erreur quadratique moyenne de prévision des filtres de lissage fondés sur l'estimation MVC.

MOTS CLÉS: Modèle linéaire; maximum de vraisemblance; maximum de vraisemblance avec contrainte; composantes de la variance.

1. INTRODUCTION

Malgré tous les efforts que l'on déploie dans un recensement pour dénombrer toute la population, les chiffres finals sont inexacts pour une multitude de raisons. Le personnel affecté au recensement – depuis le directeur de la division jusqu'aux milliers d'employés temporaires engagés comme recenseurs – est chargé d'une tâche gigantesque dont le succès dépend de la capacité de chacun d'accomplir son travail à la perfection.

En outre, il faut espérer que les manifestations de phénomènes naturels incontrôlables (par ex.: conditions atmosphériques, catastrophes naturelles) n'auront pas de conséquences inattendues sur les résultats. De toute évidence, dans un pays de la taille des E.-U. (en population comme en superficie), beaucoup de facteurs peuvent contribuer à fausser les chiffres du recensement. Cependant, la taille n'est pas le seul problème; l'hétérogénéité démographique et géographique peut fausser les chiffres de différentes façons.

Les inexactitudes s'expriment normalement par un sous-dénombrement, de sorte qu'une valeur négative signifie un surdénombrement. Supposons que les E.-U. sont divisées en i régions, $i = 1, \dots, n$, (par ex.: les 50 États plus le District de Columbia). Pour la région i , soit T_i l'effectif réel (inconnu) de la population et C_i l'effectif recensé. Alors, le sous-dénombrement, exprimé en pourcentage de l'effectif réel, est défini par l'expression:

$$U_i \equiv \{ (T_i - C_i) / T_i \} 100. \tag{1.1}$$

Les différences de taux de sous-dénombrement posent un problème sérieux lorsque les chiffres du recensement doivent servir à la répartition du pouvoir et des ressources monétaires entre les régions et les sous-régions. (Pour un examen détaillé de la question, se référer à Erickson

¹ Noel Cressie, Département de statistique, Université Iowa State, Ames, (IA), U.S.A. 50011.

Ericksen et Kadane proposent une méthode statistique inédite qui, affirmement-ils, représente une amélioration par rapport au recensement. Nous leur répondons ceci: montrez-nous. Faites-en la preuve, non selon les normes de la physique d'une part ou des recherches en matière de perception extra-sensorielle de l'autre, mais selon les normes de l'argumentation rationnelle. Au terme de deux poursuites devant les tribunaux et d'innombrables articles dans les revues spécialisées, nous concluons qu'Ericksen et Kadane n'en ont pas encore fait la preuve. Aux lecteurs d'en juger.

SOURCES ADDITIONNELLES

BERAN, R., et 12 autres statisticiens (1988). Statement on census adjustment. U.S. House of Representatives, Subcommittee on Census and Population, audition du 3 mars.

BRYANT, B. (1992). Note de service à Michael Darby et Mark Plant. Réimprimé dans *Dividing the Dollars*, un rapport du U.S. Senate Committee on Government Affairs, S Prt 102-83, Washington, DC, 78-86.

CRESSIE, N. (1987). Commentaire. *Journal of the American Statistical Association*, 82, 980-983.

FAY, R.E. (1992). Inferences for small domain estimates from the 1990 Post Enumeration Survey. Rapport technique, Bureau of the Census, Washington, DC.

FREEDMAN, D.A., KLEIN, S.P., SACKS, J., SMYTH C.A., et EVERETT, C.G. (1991). Ecological regression and voting rights. *Evaluation Review* 15, 673-711.

KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. Rapport technique, Bureau of the Census, Washington, DC.

MADANSKY, A. (1986). Commentaire. *Statistical Science*, 1, 28-30.

MULRY, M. (1991). 1990 Post Enumeration Survey Evaluation Project P16. Total Error in PES Estimates for Evaluation Post Strata. Rapport technique, U.S. Bureau of the Census, Washington, DC.

U.S. DEPARTMENT OF COMMERCE (1991a). Office of the Secretary. Decision on Whether or Not a Statistical Adjustment of the 1990 Decennial Census of Population Should Be Made for Coverage Deficiencies Resulting in an Overcount or Undercount of the Population; Explanation. Rapport daté 15 juillet, Washington, DC.

U.S. DEPARTMENT OF COMMERCE (1991b). Bulletin de presse CB91-221, daté 6/13/91.

U.S. DEPARTMENT OF COMMERCE (1991c). Bulletin de presse CB91-222, daté 6/13/91.

WOLTMAN, H.F., et coll. (1991). Loss function evaluation. Rapport technique de projet P16, U.S. Bureau of the Census, Washington, DC.

Tableau D

Simulation pour le choix de variables. Série PEP 2-9 considérée comme la réalité; la "proportion de population urbaine" (Urb) est admise parmi les variables explicatives. Le tableau indique le nombre de fois qu'a été choisie chaque variable et la valeur moyenne du coefficient correspondant (pour le nombre de fois qu'a été choisie la variable); 100 séries de données ont été produites. Dans un régime comme dans l'autre, les coefficients doivent être significatifs; selon le régime des MCG, les valeurs négatives sont permises.

Moindres carrés ordinaires		Moindres carrés généralisés	
Nombre de fois sélectionnée	Valeur moyenne du coefficient	Nombre de fois sélectionnée	Valeur moyenne du coefficient
VC	17	34	2,922
Min	82	92	0,084
Crim	53	40	0,055
Class	93	94	0,028
Ed	5	11	-0,099
Pauv	1	25	-0,212
Lang	17	5	0,417
LM	0	18	-0,048
Urb	23	19	0,053

Notes: VC est l'indicateur de ville centrale; Min, le pourcentage de minorités; Crim, le taux de criminalité; Class, la proportion de la population qui a été recensée selon la méthode classiques; Ed, la proportion de personnes qui n'ont pas de diplôme d'études secondaires; Pauv, le pourcentage de la population qui vit sous le seuil de la pauvreté; Lang, la proportion de personnes qui maîtrisent difficilement la langue anglaise; LM, le pourcentage de la population qui habite un immeuble à logements multiples.

Passons maintenant au tableau 7 et refaisons-en les calculs au moyen des MCG. Comme le montre le tableau C, le pourcentage de la population urbaine constitue une meilleure variable que le taux de criminalité, qu'on utilise les MCG ou les MCO; de plus, la série PEP 10-8 s'avère meilleure que la PEP 2-9. Les motifs invoqués par EK pour exclure la série 10-8 ne résistent pas à l'inspection.

La simulation du tableau 9 donne à peu près les mêmes résultats, que l'on retienne les variables ajustées par les MCO ou celles ajustées par les MCG. Le tableau D répète la simulation du tableau 10, en incorporant l'ajustement par les MCG. EK ont raison: urb intervient un peu moins souvent et VC, nettement plus souvent. Toujours est-il que urb s'avère supérieur à trois des variables d'EK (il s'en est fallu d'un cheveu dans le cas de LM). Par ailleurs, les signes négatifs ne sont certainement pas rares avec l'application des MCG, et ils ont les conséquences paradoxales notées ci-dessus.

Conclusions

EK affirment ce qui suit (p. 64):

Freedman et Navidi placent les données redressées à un niveau plus élevé que les données non redressées. Ils tiennent pour acquis, malgré des dizaines d'années de compilation de renseignements au Censur Bureau, que les données non redressées sont exactes et ils ne semblent pas être conscients de l'existence d'un biais important dans les régions.

Tout cela est faux. Notre article commence par une discussion des erreurs du recensement, de leur variation selon la région et des incidences sur les ressources. Nous estimons par contre que la réalisation des recensements de 1980 et 1990, dont l'exactitude globale est estimée à 98% ou 99%, constitue tout un exploit. Deux siècles d'expérience ont permis de perfectionner les compétences gestionnelles, et des centaines de milliers de citoyens ont mis la main à la pâte. Ces deux recensements n'étaient pas parfaits, mais ils ont été de francs succès.

La meilleure équation qui satisfasse les critères actuels d'EK est effectivement l'équation (MCG 12). Elle ne comporte aucun terme d'ordonnée à l'origine. Lorsque pareil terme est nécessaire, la meilleure équation est la suivante:

$$\text{PEP 2-9} = 1.260 + 2.609 \text{ VC} + .109 \text{ min} + .0262 \text{ class} - .190 \text{ pauv} + \text{résidu}$$

(2.1)

(2.9)

(5.1)

(4.1)

(-3.1)

é.q.m. = 1.56.

(VC est l'indicateur de ville centrale.) Ainsi, EK ne peuvent pas avoir sélectionné leurs variables exactement comme ils l'affirment.

Encore une fois, le critère pauv est fortement négatif. Au sein d'une ville centrale, il n'y a que deux variables: min et pauv. D'après l'équation, plus un quartier minoritaire est pauvre, plus il est facile à dénombrer.

L'équation (2) d'EKT représente une régression différente par les MCG, la matrice de covariances étant $s^2I + K$ plutôt que K , s^2 étant la valeur estimative de σ^2 et K étant la matrice de covariances fondée sur l'échantillon des sous-dénombrements bruts. Voir les équations (1-6) de notre étude.)

Nous voulions justement souligner (p. 16) qu'EK ne pouvaient pas déduire le modèle à partir des données; le passage aux MCG ne les aide pas vraiment. D'autre part, EK affirment ce qui suit (p. 63):

La vraie question est de savoir dans quelle mesure les estimations calculées à l'aide des différentes équations de régression varient entre elles. La réponse, comme nous l'avons vu plus haut, est que les taux de sous-dénombrement ne diffèrent pas sensiblement les uns des autres.

Ils soulignent ainsi un enjeu réel parmi tant d'autres (il faut notamment tenir compte de l'incidence de la sélection des variables sur les variances nominales; voir Fay 1992). Par contre, si EK reviennent à leur position de 1986, selon laquelle ils peuvent procéder à des redressements pour les sous-régions, la sélection des variables aura une importance incontestable:

Pour les 66 régions à l'étude, le choix des variables a une incidence sur le redressement, mais l'impact n'est pas appréciable étant donné que les deux jeux de variables s'étendent pour l'essentiel sur la même largeur de colonnes. Par contre, si l'on extrapole pour englober les sous-régions, le choix des variables revêt une importance considérable. (TRADUCTION) (FN, p. 9)

Tableau C

Écarts quadratiques moyens calculés à partir des équations de régression pour les séries PEP 2-9 et PEP 10-8. Les variables explicatives sont la proportion de minorités

Moindres carrés ordinaires		Moindres carrés généralisés	
Taux de criminalité	Proportion de population urbaine	Taux de criminalité	Proportion de population urbaine
PEP 2-9	1.53	1.60	1.57
PEP 10-8	1.35	1.33	1.35

Notre valeur est 7,12. (L'écart le plus important que nous avons trouvé concernait Dallas: 6,22 contre 6,18.) Armés des variables, nous pouvons nous occuper du reste.

La plus grande partie de la discussion dans FN et dans le présent document porte sur ce qu'on fait après avoir retenu les variables et n'est donc pas assujettie aux critiques d'EK concernant des erreurs de reproduction. Nous n'avons notamment pas à rougir de notre tableau 8, même si EK l'ont nommément pris à partie dans leur exposé. Ce tableau n'a en effet rien à voir avec l'algorithme pour la sélection des variables, et nous en défendons toujours le contenu. Il en va autrement de nos équations (11-12), du tableau 7 et des tableaux 9-10, qui correspondent aux tableaux 5 et 6 dans FN. Ces calculs dépendent effectivement de l'algorithme de sélection des variables, et nous en abordons l'incidence un peu plus loin, après avoir établi un contexte.

En 1986, EK ont critiqué nos simulations, mais pour d'autres motifs: les simulations s'inspiraient de la fameuse série 10-8. L'opposition des MCO et des MCG a été passée sous silence. En 1989, EKT ont critiqué les simulations de nouveau, pour d'autres motifs encore: i) nous nous sommes limités aux modèles comportant trois variables; ii) nous n'avons pas exigé que les coefficients soient significatifs.

Aujourd'hui, EK soulèvent pour la première fois la question de l'opposition des MCO et des MCG. En réponse, nous reprenons encore une fois nos calculs, en faisant appel aux MCG et à des observations "pondérées par l'inverse de l'erreur-type des estimations d'échantillon initiales"; les coefficients doivent être significatifs, mais les valeurs négatives sont permises. Nous traitons d'abord des équations (11) et (12); les valeurs *t* figurent entre parenthèses.

$$\begin{array}{l} \text{(MCO 11)} \quad \text{PEP 2-9} = -2,23 + ,079 \text{ min} + ,036 \text{ crim} + ,028 \text{ class} + \text{résidu} \\ \text{é.q.m.} = 1,53. \end{array}$$

$$\begin{array}{l} \text{(MCO 12)} \quad \text{PEP 2-9} = ,120 \text{ min} + ,026 \text{ crim} + ,029 \text{ class} - ,176 \text{ pauv} + \text{résidu} \\ \text{é.q.m.} = 1,49. \end{array}$$

$$\begin{array}{l} \text{(MCG 11)} \quad \text{PEP 2-9} = -3,37 + ,054 \text{ min} + ,061 \text{ crim} + ,026 \text{ class} + \text{résidu} \\ \text{é.q.m.} = 1,60. \end{array}$$

$$\begin{array}{l} \text{(MCG 12)} \quad \text{PEP 2-9} = ,118 \text{ min} + ,030 \text{ crim} + ,031 \text{ class} - ,217 \text{ pauv} + \text{résidu} \\ \text{é.q.m.} = 1,53. \end{array}$$

Min est le pourcentage de minorités; crim, le taux de criminalité; class, la proportion de la population qui a été recensée selon la méthode classique; pauv, le pourcentage de la population qui vit sous le seuil de la pauvreté. Ainsi qu'on pourra le voir, les pondérations ont une incidence qualitative négligeable (bien que l'écart soit appréciable au niveau des valeurs *t*). Sous un régime comme sous l'autre, la valeur de pauv est très significative. En outre, l'équation qui tient compte de la pauv est supérieure, car les résidus y sont moins élevés.

Plus une région est pauvre, *moins* le sous-dénombrement y est important. C'est du moins ce que "montre" l'équation (12); les autres variables (par exemple, la composition raciale, le taux de criminalité et la méthode de dénombrement) sont contrôlés au moyen de la régression. Cette constatation ne concorde pas avec la théorie du sous-dénombrement d'EK, malgré l'argument ingénieux des auteurs à la page 62.

De même, EK affirme (p. 58) que «la méthode synthétique B de Schirm et Preston, [...] même si elle représente une amélioration par rapport aux chiffres non redressés, est nettement insuffisante». Cela contredit les positions précédentes du panel, qui étaient cependant provisoires; voir Cohen et Citro (1985, p. 287) et notre étude (p. 8).

Mise en moyenne et analyse de sensibilité

EK nous invite (p. 60) à remplacer les diverses séries du PEP par la moyenne et à considérer les écarts de l'é.q.m. par rapport à la moyenne. En établissant douze formules d'imputation différentes, on cherchait cependant à mesurer l'impact de la modélisation. C'est pourquoi l'intervalle de variation constitue la statistique appropriée: deux modèles d'imputation choisis au hasard peuvent donner des résultats semblables, mais un troisième peut être fort différent. Au bout du compte, EK veut procéder à une analyse de sensibilité, mais ils minimisent l'importance de tout modèle qui se distingue des autres.

EK propose encore de sous-échantillonner, des estimations de chaque série pour les 66 régions étudiées, le taux de sous-dénombrement estimatif à l'échelle nationale. Ils supposent tacitement – et sans fondement – qu'un modèle d'imputation vaut pour l'ensemble du pays. Notre analyse suppose pour sa part que différents motifs peuvent expliquer les données manquantes dans diverses parties du pays (sec. 7.1).

D'autre part, EK propose de nouveaux motifs pour rejeter la série PEP 10-8 (p. 61). L'argument se résume comme suit: si les séries qu'ils jugent supérieures sont valables, notre série-repère 10-8 ne l'est pas. Exactement. Inversement, si la série 10-8 est valable, leurs séries ne le sont pas. Autrement dit, le choix de la série du PEP à utiliser a son importance. Lorsqu'on estime de faibles taux de sous-dénombrement, des données manquantes qui représentent huit points de pourcentage comptent pour beaucoup. Aucune manipulation statistique ne peut modifier cette réalité gênante.

Reproduction

EK soutiennent ce qui suit:

L'inconvénient de la critique de FN est qu'ils n'ont pas reproduit correctement notre méthode de sélection. Comme nous l'avons expliqué dans EKT et dans d'autres ouvrages (Ericksen et Kadane 1985; 1987, section 6),

- i) les observations ont été pondérées par l'inverse de l'erreur-type des estimations d'échantillon initiales.
- ii) «[...] L'estimation du taux de sous-dénombrement que nous avons calculée est la moyenne pondérée d'une estimation par régression et des estimations d'échantillon initiales [...]» [p. 61-62; l'ordre des points a été interverti]

Malheureusement, EK confondent deux questions: i) comment choisir des variables; ii) quoi faire après les avoir choisies. S'agissant du point ii), EK utilisent le modèle de régression bayésien hiérarchique de Lindley-Smith. Il n'y a qu'une ombre au tableau: le paramètre σ^2 , inconnu, doit être estimé; voir FN, p. 5 et 11.

Une fois les variables sélectionnées et σ^2 estimé, il ne reste aucune ambiguïté quant à l'estimateur d'EK. Voir l'équation (3) d'Ericksen-Kadane (1985), l'équation (9) de FN ou les notes au tableau 8 de notre étude principale. En effet, nous avons pu reproduire leurs chiffres devant le tribunal, le juge a même souligné l'exactitude des ordinateurs de Berkeley. Illustrons le point à nouveau, au moyen de données tirées d'EKT. On peut extraire du tableau 10 leur estimateur composite fondé sur la série PEP 2-9. Leur premier exemple est la ville de Saint Louis, et la valeur qu'ils obtiennent pour l'estimateur est la suivante:

$$1.24 + 0.66 + 4.16 + 1.09 = 7.13.$$

EK donne à entendre (p. 59) que la concordance compte pour beaucoup, puisque l'analyse démographique "produit une estimation juste du taux de sous-dénombrement national". L'analyse démographique est vraisemblablement plus fiable que le PEP. Elle n'est pas pour autant exempte de vrais problèmes; voir la discussion sur l'article de Cressie ci-dessus. La concordance constitue un argument faible.

Par ailleurs, toute correspondance entre le PEP et l'analyse démographique au niveau de l'ensemble masque des écarts appréciables au niveau désagregé, comme Jeff Passel l'a prouvé devant le tribunal. Nous avons résumé les arguments ailleurs, mais nous les reprenons encore une fois. La série PEP 2-9 est considérée comme la *nec plus ultra* des séries jugées supérieures par EKT, le tableau B compare les résultats de la série 2-9 à l'analyse démographique. Ainsi, la série PEP 2-9 sous-estime légèrement les hommes de race noire et surestime de 100% les femmes de race noire, alors qu'elle sous-estime de 33% les hommes de race blanche et surestime les femmes de race blanche (0,5% contre 0). La correspondance s'est dissipée.

Par surcroît, la défense d'EK n'est pas tout à fait cohérente. Comparons, par exemple, les pages 59 et 60. À la page 59, les taux de sous-dénombrement à l'échelle nationale sont importants, alors qu'à la page suivante, la variation des taux de sous-dénombrement à l'échelle nationale a une importance négligeable. De plus, la position du moment – mise en moyenne sur un certain nombre de pages – ne concorde pas avec celle prise par le National Academy of Science Panel, où siégeaient notamment Steve Fienberg et Jay Kadane:

Il existe plusieurs motifs, tant *a priori* qu'*a posteriori*, pour appuyer les diverses séries PEP individuelles tirées de cette liste de douze. . . Par exemple, l'estimation 10-8 atténue le problème des personnes ayant déménagé du fait qu'elle utilise l'échantillon *P* du mois d'août. . . On retrouve ces points, et d'autres encore, dans l'affidavit de Bailar relatif à la cause *Cuomo v. Baldrige*. (TRADUCTION)

Le recours à ces douze estimations a produit des estimations fort différentes de la sous-représentation des groupes démographiques à l'échelle nationale. . . Certains analystes proposent de rétrécir considérablement le nombre d'estimations 2-8, 2-9, 3-8 et 3-9, parce qu'elles sont fondées sur des données du mois d'août (où il y a un taux plus élevé de cas d'appariement non réglés) ou parce qu'elles s'appuient sur des hypothèses extrêmes pour redresser les données manquantes. Même au sein de cette série restreinte, toutefois, le taux de sous-dénombrement à l'échelle nationale varie entre 0.8% et 1.4%. [Cohen et Citro 1985, p. 147-148]

Bref, même parmi les séries jugées supérieures par EK, les résultats varient selon le modèle d'imputation utilisé. Il n'existe également aucun motif valable d'établir une discrimination contre notre série-repère 10-8.

Tableau B

Comparaison de deux méthodes d'estimation des taux de sous-dénombrement dans le recensement de 1980: analyse démographique (AD) et série PEP 2-9.

	AD	PEP 2-9
Hommes de race noire	8.8	8.1
Femmes de race noire	3.1	6.4
Hommes de race blanche	1.5	1.0
Femmes de race blanche	0.0	0.5

Source: Fay et coll. 1988, annexe D.
Note: L'analyse démographique est fondée sur la série DA-2; l'expression "race blanche" englobe les "autres races".

modèle de régression. Pour amorcer leur modèle, EK s'en tiennent à une phrase: "Il n'y a rien qui laisse croire que la combinaison de sources d'information soit une opération incohérente ou inhabituelle". La faiblesse de l'argumentation surprend, car les seuls critères à respecter sont les suivants: i) le modèle ne devrait comporter aucune contradiction interne; ii) un tiers devrait déjà avoir fait quelque chose de semblable.

Le redressement pour les petites régions

Plus tôt, EKT ont semble concéder qu'ils ne pouvaient pas procéder à un redressement pour les petites régions (p. 943, ainsi que la section 4 de notre article principal). EK se ravisent maintenant (p. 60), citant les travaux de Tukey et de Wolter et Causey. Nous avons passé ces travaux en revue dans l'annexe de notre étude. Nous ne les avons pas trouvés convaincants et avons expliqué pourquoi. EK ne répondent pas à nos arguments.

Le tableau 5 d'EKT

EK prétendent que notre argumentation va trop loin. Pourtant, leur tableau 5 est censé montrer que les séries du PEP qu'ils jugent supérieures correspondent, en règle générale, aux estimations synthétiques. Une telle correspondance ne démontrerait la valeur du PEP que si on savait les estimations synthétiques exactes. Comme on l'a vu plus haut, cette prémisse est douteuse.

De plus, étant donné l'échelle retenue par EKT, nous avons trouvé un niveau remarquable de discordance parmi les séries du PEP qu'ils jugent supérieures. La réponse d'EK: après avoir privilégié huit séries du PEP sur douze, ils veulent maintenant en éliminer deux autres (du mois d'août). La démarche est mauvaise, notamment du fait que l'écart extrême noté dans notre équation (8) tient compte des séries du mois d'avril. Ensuite, EK établissent une moyenne pour leurs séries jugées supérieures. La mise en moyenne des résultats d'une analyse de sensibilité dans le but de réduire la variation représente une notion bizarre, comme nous l'avons expliqué dans la section 5 de notre étude principale. Nous y reviendrons plus loin.

EK poursuivent comme suit (p. 59):

Il y a un point plus important encore toutefois: les quatorze formes de redressement [soit les douze séries du PEP et les deux redressements synthétiques] ont pour effet d'améliorer les chiffres du recensement en transférant des parts de population des régions où il n'y a pas eu beaucoup de problèmes de recensement vers celles où les problèmes étaient nombreux.

Cela tient de l'euphémisme. Comme en atteste le tableau 5 d'EKT (voir aussi EKT, p. 927 et EK, p. 58), les régions où les problèmes de recensement étaient nombreux corresponaient à celles comptant une forte concentration de personnes appartenant à une minorité. Pourtant, comme nous l'avons expliqué en répondant à Fienberg, les sièges de la députaion et les recettes fiscales sont répartis parmi des secteurs géographiques plutôt que parmi des groupes raciaux ou ethniques. La question fondamentale est de savoir si le redressement améliorerait l'exactitude des parts de population pour les petites régions, à savoir les Etats, les villes et les comtés. Le tableau 5 d'EKT, pour sa part, concerne de vastes regroupements de villes et d'Etats. Ces groupements paraissent artificiels.

Quelle série du PEP faut-il utiliser?

EK tentent encore une fois (p. 59) de justifier la supériorité qu'ils attribuent à huit des douze séries du PEP en lice: ils cherchent particulièrement à éliminer notre série-repère 10-8. Ils invoquent, pour argument principal, la concordance avec l'analyse démographique à l'échelle nationale. Ils soutiennent également que "les résultats relatifs aux régions à forte concentration de Noirs sont conformes aux résultats de l'analyse démographique". Il doit y avoir eu un glissement de sens, car l'analyse démographique ne donne aucun résultat en deçà du niveau national.

L'argument qu'ils invoquent se retrouve à la page 60:

Freedman et Navidi prétendent que certaines des hypothèses qui sous-tendent notre modèle de régression sont "non vérifiées" et "peu plausibles". Comme nous l'avons déjà soutenu, que ce soit dans EKT ou dans d'autres ouvrages (Ericksen 1986; Kadane 1986), nous croyons que nos hypothèses sont réalistes et qu'elles reposent sur un bloc de connaissances accumulées depuis des décennies.

Nous avons passé en revue l'exposé d'EKT et les deux ouvrages cités (Ericksen 1986 et Kadane 1986). Nous n'avons trouvé aucune preuve empirique susceptible de justifier les hypothèses, de quantifier les échecs (par exemple, de déterminer la taille réelle des corrélations établies par hypothèse à 0) ou de déterminer l'impact des échecs sur les résultats des modèles. EK se tient plutôt à des arguments de commodité (le modèle est "simple et résolvable" et il "permet le lissage", Kadane 1986, p. 13). Ils ont également mis au point leur propre variation sur le thème "rien n'est parfait":

[...] dans les applications, seul un utilisateur fort naïf croirait à la vérité littéraire des hypothèses. Ainsi, lorsque je propose et utilise une hypothèse, je veux dire que je crois à la vérité de quelque chose qui s'en rapproche, mais je ne veux sûrement pas dire que c'est exactement ça la vérité [...] (TRADUCTION) (Kadane 1986, p. 14).

Qu'est-ce qui incite EK à penser que "quelque chose qui se rapproche" de leur modèle est vrai? Même lorsqu'elles sont réunies, la commodité et l'imperfection de ce monde ne valident pas une hypothèse et ne quantifient pas l'impact d'un échec.

Passant à un autre front, EK nous accusent d'avoir posé nos propres hypothèses non vérifiées. Notre culpabilité à cet égard ne nous entend certainement pas leur innocence; quoi qu'il en soit, nous nions les accusations, du moins pour la plupart. Trois exemples donnent la saveur de nos "hypothèses non vérifiées" (p. 58):

i) On relève toujours un certain niveau de sous-dénombrement, modeste, dans les recensements.

ii) Les membres de minorités qui vivent dans les villes centrales ont des chances d'être différents de ceux qui vivent dans les banlieues.

iii) Le taux de sous-dénombrement était relativement élevé dans les régions ayant fait l'objet d'un recensement classique.

L'hypothèse i) semble toujours valable. Si EK le concédaient, nous pourrions tous faire gagner beaucoup de temps aux tribunaux et laisser à d'autres collaborateurs les pages des revues spécialisées. L'hypothèse ii) est évidente pour quiconque a passé quelques jours dans une grande ville américaine; s'il faut absolument des données, toutefois, voir Freedman et coll. (1991), où l'on fait aussi le point sur certains ouvrages connexes. Quant à l'hypothèse iii), nous aurions dû préciser "taux de sous-dénombrement estimatif". Touché!

La preuve des hypothèses

Un exemple suffit. EK affirment (p. 57) ce qui suit:

[...] le recensement s'effectue plus difficilement dans certaines régions et [...] les taux de sous-dénombrement calculés à l'aide des données du PEP sont plus élevés dans les régions où le taux de réponse par la poste est relativement faible et la proportion de données manquantes plus élevée et où il est plus difficile de respecter le taux d'échantillonnage spécifié pour les questionnaires complets du recensement.

(Pour un examen sommaire des séries du PEP, voir les sections 2 et 8 ci-dessus.) Ici, EK prouvent tout au plus que les données du PEP ont un rapport avec les taux de sous-dénombrement; nous ne l'avons jamais nié. Il n'est cependant pas possible de résumer tous les rapports dans un

non due à l'échantillonnage. (De plus, les deux modèles exercent l'un sur l'autre des effets importants; nous proposons d'y revenir un autre jour.) Les partisans du redressement cherchent à établir un sous-dénombrement d'environ 2%. Pour y arriver, ils doivent maintenir bien inférieure à 1% l'erreur non due à l'échantillonnage de l'EP. Ils affirment y avoir réussi, à partir des données d'un autre sondage, le suivi d'évaluation. S'ils mesurent les erreurs non dues à l'échantillonnage de l'EP à une fraction de 1% près, les erreurs du suivi d'évaluation doivent être encore plus petites. Nous fait-on marcher?

Les cinq questions

Hartigan conclut en posant cinq questions. Nous répondrons à deux d'entre elles. (La première a été légèrement remaniée, pour plus de clarté.)

i) "Freedman et Navidi sont-ils d'accord avec ces estimations de 17 millions de personnes oubliées et de 13 millions d'enregistrements erronés?" Nous les acceptons comme des approximations, assujetties à des biais et à des erreurs-types importants et inconnus. L'écart de $17 - 13 = 4$ millions risque d'être fautif par un facteur de 2 ou plus. Estimer un petit nombre en prenant la différence entre deux grands, voilà un bon moyen de s'attirer des ennuis. Par ailleurs, il reste la question épineuse de savoir où situer les 4 ± 2 millions de personnes. Fienberg n'aime pas les collines du Dakota du Sud. Il reste donc 6,5 millions d'îlots répartis sur 39,000 territoires. L'EP nous a fourni des données sur 0,2% des îlots et sur peut-être 10% des territoires. Un chef-d'œuvre théâtral oblige la salle à suspendre son incrédulité. Le redressement n'est manifestement pas à la hauteur.

ii) "Si l'EP ne donne pas des résultats satisfaisants, comment l'enquête de suivi devrait-elle être conçue afin qu'on puisse en utiliser les résultats pour redresser les chiffres du recensement?" Nous répondons à Hartigan par une autre question: qu'est-ce qui lui fait penser que cette entreprise est le moins réalisable?

12. Speed

Le redressement dépend de modèles et d'hypothèses pour lesquelles il n'existe aucune preuve empirique. Tel est l'avis de Speed, que nous partageons.

Afin de redresser le recensement de 1990, on répartit la population parmi 1,392 groupes démographiques ou "strates *a posteriori*". La strate *a posteriori* 90302112, par exemple, regroupe les locataires masculins d'origine hispanique âgés de 10 à 19 ans dans les villes de la division du Pacifique. Le redressement repose sur l'"hypothèse d'homogénéité", voulant que les taux de sous-dénombrement soient plus ou moins constants pour chacune des strates *a posteriori* des diverses régions. Voir Freedman (1991) ou U.S. Department of Commerce (1991a, p. 2.37-2.45, p. 4.16-4.18).

Cette hypothèse n'a certainement rien d'une réalité évidente. Le Bureau a procédé à une étude pour la mettre à l'épreuve (Kim 1991). L'étude en question semble cependant avoir été très mal conçue et elle ne donne de toute façon que des résultats mitigés. La théorie du redressement est particulièrement fragile lorsqu'on l'applique aux petites régions.

13. Ericksen et Kadane

Le rôle des hypothèses

Nous affirmions que "l'efficacité de l'un ou l'autre des redressements proposés par EKT repose sur des hypothèses non vérifiées et peu plausibles". EK répondent (page 58) que leurs "hypothèses sont réalistes et qu'elles sont vérifiées par des dizaines d'années d'expérience en recensements, comme nous le verrons plus loin."

ou par enquête postcensitaire, en deçà d'un facteur de 2. Voir, ci-dessus, la discussion sur l'article de Cressie. En outre, Hartigan ne tient pas compte des variations dans le taux de sous-dénombrement des non-minorités entre les Etats. Il faut cependant que ces variations aient une importance: par exemple, un sous-dénombrement de 1% dans un Etat qui compte 9 millions de personnes a presque deux fois l'impact d'un sous-dénombrement de 5% dans un Etat qui en compte 1 million.

La modélisation

Hartigan écrit aussi ce qui suit:

Je soupçonne que les hypothèses utilisées pour la régression ne peuvent être défendues facilement, mais que les résultats de la régression sont raisonnables, sauf peut-être qu'ils produisent des erreurs-types plus faibles que le manque probable d'indépendance ne le justifie. . . La réduction de la variance d'échantillonnage à l'aide de procédures de lissage basées sur la régression ne fera probablement pas beaucoup de différence pour les estimations dans de grandes régions comme les Etats. . . On est justifié de faire preuve de scepticisme de bon aloi à propos de toutes les "erreurs-types" ou de tous les "intervalles de confiance" résultants. [p. 51, soulignement omis]

En ce qui concerne les données de 1980, le choix des variables a une incidence appréciable sur le redressement dans les petites régions. Voir FN, page 9. S'agissant des données de 1990, les facteurs de redressement "bruts" (calculés directement à partir de l'échantillon, sans régression) sont entachés d'erreurs d'échantillonnage tellement élevées qu'ils sont inutilisables, même au niveau des Etats. Aussi les partisans du redressement doivent-ils faire appel au lissage. Par contre, le choix du modèle de lissage exerce une influence déterminante sur les résultats. Voir la décision du Secrétaire (U.S. Department of Commerce 1991a, p. 2.46-2.55) et les chiffres donnés dans le communiqué (U.S. Department of Commerce 1991b).

Par ailleurs, l'argument en faveur du redressement s'appuie sur une "analyse de la fonction de perte", laquelle fait appel à des variances calculées à partir du modèle de lissage pour procéder à des estimations non biaisées du risque. On sait toutefois que le modèle est trop optimiste quant aux variances, peut-être par un facteur de 5; voir FN, p. 10, Yivisaker (1991), notre étude principale sec. 7.4 et Fay (1992). Si on fait preuve de "scepticisme de bon aloi" à l'égard des fonctions de perte, il ne reste à notre avis aucun argument sur la table en faveur de l'efficacité des redressements proposés.

Nous entendons nous pencher sur l'analyse de la fonction de perte du Bureau dans une autre étude. Hartigan propose ses propres calculs aux pages 53 et suivantes; encore une fois, ils s'écartent trop des données pour qu'on les prenne au sérieux. De toute façon, le lecteur peut consulter l'analyse du Bureau (Mulry 1991; Wolman et coll. 1991) avant d'accepter quelque conclusion que ce soit.

La troisième vérité

Le Bureau a produit un certain nombre de sous-dénombrements estimés, avec des marges d'erreur, dans les divers Etats. J'utilise "la méthode choisie pour l'EP" (que je désignerai dorénavant par l'abréviation EP). . . Nous possédons maintenant deux vérités, le dénombrement et les chiffres de l'EP. Laquelle est exacte? Bien, il nous faut une troisième vérité, une qui ne sera pas contestée. . . [p. 52]

Hartigan touche ici à quelque chose de capital. La "troisième vérité" du Bureau consiste en l'analyse de fonction de perte discutée ci-dessus et en un "modèle de l'erreur totale" (Mulry 1991). Ces deux éléments semblent très contestables: l'analyse de fonction de perte parce qu'elle s'appuie sur des variances calculées à partir du modèle de lissage, et le modèle de l'erreur totale parce qu'il se fonde sur les résultats du suivi d'évaluation pour mesurer l'erreur

Adoptons brièvement leur terminologie. Ils ont construit des populations réelles à partir de l'hypothèse de la méthode synthétique, à laquelle s'ajoute l'erreur aléatoire. En effet, les simulations gardent les chiffres du recensement fixes et rendent aléatoire le chiffre de la population réelle. La population réelle de la race j dans l'Etat i est supposée égale au chiffre du recensement correspondant, multiplié par un facteur de redressement aléatoire u_{ij} . Voir SP (1987), équation (1) à la page 967. Ce facteur de redressement est tiré au hasard d'une distribution qui dépend, par hypothèse, du groupe racial, mais non de l'Etat. Voir SP (1987), équation (2) à la page 967.

Les simulations partent de l'hypothèse qu'il n'y a aucune variation systématique du taux de sous-dénombrement pour une race donnée d'une région à une autre. Par contre, le redressement synthétique suppose que la structure du sous-dénombrement est déterminée par la race, et non par la région. C'est justement ce que nous disions à la page 11, et c'est vrai.

S et P le concèdent effectivement (p. 40):

Nous avons considéré des cas extrêmes, bien que non systématiques, de variation entre les Etats du taux de sous-dénombrement selon la race. . .

L'expression "bien que non systématiques" représente leur concession; le mot "extrêmes" doit être leur défense.

L'argument *a fortiori*

S et P affirment que leurs simulations étaient prudentes, car la structure réelle de la variation des taux de sous-dénombrement par région favoriserait encore davantage le redressement synthétique que les hypothèses qu'ils ont posées. (Voir, par exemple, page 41.) SP (1987) avant-gaient des arguments *a priori* à cet effet. Passel (1987) montre notamment que de tels arguments ne prouvent pas grand-chose en ce qui concerne 1980; voir l'annexe à notre étude. SP (1987, p. 977) proposent des arguments empiriques, fondés sur des données qui "renferment beaucoup de lacunes et [sont] basées sur des hypothèses fragiles"; ce sont les termes de S et P (p. 41), et non les nôtres. Il ne semble pas nécessaire de poursuivre le débat.

Aux pages 40 et 41, S et P introduisent une nouvelle analyse fondée sur le PEP afin de justifier les paramètres de leurs simulations. Il s'agit, dans ce contexte, d'une stratégie assez surprenante: EKT s'attendent à ce qu'on se fie aux séries du PEP parce qu'elles ressemblent aux redressements synthétiques, alors que S et P voudraient qu'on se fie aux redressements synthétiques parce que les simulations ressemblent aux séries du PEP.

Avant d'accepter l'un ou l'autre de ces arguments, nous voulons des preuves. Le raisonnement circulaire n'est guère persuasif.

11. Hartigan

Le redressement synthétique

Hartigan rejette Schirm et Preston, mais plaide vigoureusement en faveur du redressement synthétique (pages 48 et 49). Il déclare notamment ce qui suit:

Que faut-il penser de l'argument analytique suivant? Supposons que les sous-dénombrements nationaux soient estimés correctement, mais qu'ils diffèrent dans les Etats. . . Les taux de sous-dénombrement à l'échelle nationale de 5% et de 1% sont appuyés par des données historiques du Bureau, obtenues tant à la suite d'analyses démographiques que d'enquêtes postcensitaires. . . Je ne tiendrai pas compte des variations dans le sous-dénombrement des non-minorités entre les Etats. . .

Sauf erreur monumentale, ces arguments analytiques s'éloignent trop des faits pour être pertinents. L'hypothèse fondamentale de Hartigan, c'est que les taux de sous-dénombrement à l'échelle nationale sont connus. Cette hypothèse est parfois fausse; nous ne croyons pas qu'il soit possible d'estimer les taux avec la moindre fiabilité, que ce soit par analyse démographique

tiers et les deux tiers de ces 2,1%. Voir Mulry (1991, tableau 15) et U.S. Department of Commerce (1991d). L'EP semble entachée d'un défaut fatal. Nous y reviendrons plus loin, au moment de répondre à Hartigan.

Conclusion

La thèse principale de Cressie serait la suivante (p. 35):

Pour résoudre un problème aussi difficile que le redressement des chiffres en fonction du sous-dénombrement, il faut reconnaître le but commun. Une fois que cela est fait, la discussion devrait se concentrer sur les différences dans les moyens qui pourraient permettre d'atteindre ce but. Si Freedman et Navidi pensent que cela est impossible (ce qu'ils semblent avoir laissé supposer au fil des ans), alors il faudrait le dire explicitement. Disons les choses telles qu'elles sont. A notre avis, le PEP ne pouvait pas venir à bout du problème en 1980, et l'EP ne peut pas en venir à bout en 1990. Et nous ne débordons pas d'optimisme quant aux perspectives pour l'an 2000, quel que soit l'acronyme utilisé à ce moment-là. Lorsqu'on n'arrive pas à dénombrer des personnes, il faut éviter d'inventer des chiffres par la suite en faisant passer des données de saisie-resaisie par des modèles de lissage.

9. Passel (1987)

Plusieurs critiques défendent le redressement synthétique, fort vigoureusement dans certains cas. Notre contre-exemple ne fait pour sa part que de rares adeptes (tableau 3). Néanmoins, Passel (1987) s'est servi de données du recensement de 1980 pour montrer que le redressement synthétique était peu susceptible d'améliorer l'exactitude. Nous avons résumé son travail dans l'annexe à notre étude. Aucun critique n'a répondu à son argument.

10. Schirm et Preston

Le contre-exemple

SP (1987, p. 966) prétend au redressement synthétique la propriété suivante:

Nous avons conclu que le redressement synthétique rapproche toujours le ratio estimatif et le ratio réel de la population d'un Etat à la population nationale lorsque a) le sous-dénombrement des Noirs dans l'Etat se rapproche davantage du sous-dénombrement des Noirs à l'échelle nationale que du sous-dénombrement des deux races réunies à l'échelle nationale, et que b) le sous-dénombrement des Blancs dans l'Etat se rapproche davantage du sous-dénombrement des Blancs à l'échelle nationale que du sous-dénombrement des deux races réunies à l'échelle nationale. (TRADUCTION)

Notre contre-exemple (tableau 3) a montré que ce résultat était faux. Schirm et Preston devraient le concéder.

Dans certaines conditions et selon certains critères, le redressement synthétique est sans doute avantageux; voir, par exemple, le lemme 15 de notre étude. Le résultat publié dans l'annexe de SP (1987) est exact, mais il n'est guère éclairant: l'inégalité que les auteurs supposent dans l'équation (A.2) à la page 976 est précisément l'inégalité de l'erreur absolue qu'ils cherchent à démontrer, à un facteur d'échelle près.

Les simulations

S et P préfèrent une définition rigoureuse de l'hypothèse synthétique: "il n'y a aucune variation" dans le taux de sous-dénombrement d'une race donnée d'une région à une autre. Ils affirment (p. 40) ne pas avoir construit de population réelle suivant l'hypothèse de la méthode synthétique et affirment que leur définition de la réalité n'a pas favorisé le redressement synthétique.

Lorsque Cressie aborde les cas, il calcule des risques estimatifs (des pertes prévues). Voir ses équations (2.28-31). C'est dire qu'il doit calculer des variances, lesquelles, comme il sait bien, sont très sensibles aux hypothèses:

Il va sans dire que ces résultats dépendent de la justesse du modèle hypothétique. (TRANSDUCTION) (p. 193)

Une illustration élémentaire s'impose peut-être. Supposons $\epsilon_1, \dots, \epsilon_{66}$ interchangeables, et une moyenne 0, une variance σ^2 et une corrélation par paire ρ . Maintenant,

$$\binom{66}{2} = 2145.$$

Par conséquent,

$$\text{var}\{\epsilon_1 + \dots + \epsilon_{66}\} = (66 + 2145\rho)\sigma^2.$$

Dans un tel contexte, une corrélation de 0.05, par exemple, a un impact énorme. Par surcroît, une corrélation aussi faible serait très difficile à déceler empiriquement. Cressie n'essaie pas. (Avec 16 points de donnée, même une corrélation de 0.5 risque d'être difficile à estimer, de sorte que le test n° 3 d'EKT, à la page 931, ne peut pas être déterminant.)

L'exemple peut sembler artificiel. Toutefois, l'erreur d'échantillonnage a constitué un obstacle d'envergure lorsqu'on a voulu redresser le recensement de 1990 sur la base de l'EP, même au niveau des États. D'ailleurs, d'après des données publiées, les parts de la population attribuables à une nette majorité d'États en vertu du redressement seraient à deux erreurs-types près des parts obtenues à partir du recensement (U.S. Department of Commerce 1991b). De tels redressements pourraient être imputables, dans leur totalité, à l'erreur d'échantillonnage de l'EP. (Les partisans du redressement répliqueront peut-être en invoquant l'"analyse de la fonction de perte", et nous envisagerons cette possibilité brièvement au moment de répondre à Hartigan.)

Les erreurs-types, comme les redressements estimatifs, sont les résultats d'un modèle de lissage semblable au modèle EK. À la lumière d'expériences d'auto-amorçage signalées dans Fay (1992), ces erreurs-types sont trop faibles par un facteur d'environ 2. (Fay fixe une fourchette allant de 1.4 à 2.2, avec un multiplicateur préféré de 1.7.) Lorsqu'il s'agit de calculer des variances, les hypothèses font toute la différence.

PEP 3-8

À la page 34, Cressie admet que les données du redressement de 1980 n'étaient pas assez fortes pour servir. À la page 36, il veut procéder au redressement en utilisant son modèle et la série PEP 3-8 (voir la section 7 ci-dessus). Il semble avoir écarté, par le simple jeu de ses hypothèses, tous les problèmes suscités par l'erreur non due à l'échantillonnage, les corrélations manquantes, etc. Ses calculs n'auraient alors aucun rapport avec les questions de principe que nous nous posons.

La qualité de l'EP

D'après Cressie (p. 36), l'enquête postcensitaire a été "bien conçue, bien appliquée et avec une bonne assurance de la qualité". Elle l'a effectivement été, comparativement à une étude de marché ordinaire ou peut-être même à d'autres enquêtes menées par le Census Bureau. Pourtant, lorsqu'on veut corriger une petite erreur dans le recensement, on a besoin d'une enquête par sondage dont les erreurs sont d'un ordre beaucoup plus petit. Nous ne croyons pas que l'EP respecte cette norme. À titre d'exemple, l'EP a estimé à 2.1% le taux de sous-dénombrement national. Or, l'erreur non due à l'échantillonnage dans l'EP explique entre le

Tableau A

Sommaire des révisions apportées à l'analyse démographique du recensement de 1980: sous-dénombrements estimatifs par date d'estimation.

	1984	1988	1991
Ensemble des races	0.5	1.4	1.2
Noirs	5.3	5.9	4.5
Non-Noirs	-0.2	0.7	0.8
Ecart	5.5	5.2	3.7

Source: Col. 1. Cressie, citant Passel et Robinson (1984); le chiffre pour l'ensemble des races est dérivé. Col. 2. Fay et coll. (1988, p. 95, série DA-2). Col. 3. U.S. Department of Commerce (1991c, tableau 3).

une année passée (1985, par exemple) dépend-il de l'année où se fait l'estimation. Les chiffres ne cessent d'être modifiés, et les modifications donnent une idée de la fiabilité des données initiales.

Le tableau A récapitule brièvement les révisions apportées à l'analyse démographique pour le recensement de 1980. Comme on le verra, les chiffres sont loin d'être stables. Le changement observé de 1984 à 1988 témoigne peut-être d'une nouvelle appréciation du rôle de l'immigration illégale. Celui observé de 1988 à 1991 peut, pour sa part, refléter l'impact des redressements apportés à des redressements préalables dans le but de compenser le sous-enregistrement des naissances au cours de la période allant de 1935 à 1960. Il s'agirait en effet de surredressements, qui ont peut-être déjà été corrigés.

Les démographes peuvent estimer la population des Etats-Unis à partir des données tirées des dossiers administratifs, avec une marge d'erreur d'à peine 1 ou 2 points de pourcentage, ce qui représente tout un exploit. Il semble cependant peu vraisemblable que l'erreur soit bien inférieure à un point de pourcentage. Si tel est le cas, l'analyse démographique risque de ne pas être suffisamment fiable pour servir au redressement du recensement.

Modélisation

Cressie affirme ce qui suit (p. 35):

Le concept important qui doit être conservé à l'esprit est que le véritable sous-dénombrement dans les régions est inconnu et que cette ignorance est quantifiée dans un modèle probabiliste. Le but n'est pas d'estimer les coefficients β mais de prédire le sous-dénombrement. Avec un terme d'erreur qui n'a pas à être indépendant et distribué de façon identique, cette prévision n'est pas sensible aux erreurs de spécification. . . . [soulignement omis]

Nous ne sommes pas d'accord. On n'appuie pas une politique d'intérêt public sur un modèle inspiré de l'ignorance d'un chercheur. Les résultats du modèle doivent d'ailleurs être reliés étroitement aux spécifications. En guise d'illustration, notons quelques-unes des hypothèses du modèle élaboré par Cressie (1988). Les équations (2.7) et (2.10) de ce document écartent effectivement l'erreur non due à l'échantillonnage dans l'EP, de même que la variation systématique des taux de sous-dénombrement d'un secteur géographique à un autre, et aucune corrélation n'apparaît. Pourquoi? L'équation (2.10) précise une variance d'échantillonnage qui évite les incohérences internes du modèle d'Ericksen-Kadane (Cressie 1988 p. 193). La cohérence logique ne sous-entend cependant pas la réalité empirique. Où intervient le véritable plan de sondage? Enfin, pourquoi devrait-on utiliser la fonction de perte de Cressie (2.15)? Tant que Cressie n'aura pas répondu à ces questions, et à d'autres semblables, il ne doit pas s'attendre à ce qu'on prenne sérieusement les résultats de son modèle.

En effet, nous nous attendions à ce que les modélisateurs proposent une argumentation sérieuse quant à la validité des hypothèses, plutôt que des intuitions quant aux sources éventuelles de dépendance telles que les tempêtes de neige et les volcans. Avec le temps, nos attentes se sont atténuées. D'un côté comme de l'autre, il s'avère difficile d'obtenir de vraies preuves empiriques. Pour les partisans du redressement, les modèles se justifient par la notion suivante: rien n'est parfait, de sorte qu'on peut faire n'importe quoi. En revanche, le recensement est tenu d'être exact à quelques points de pourcentage près, lorsque l'exacitude est définie par les modèles.

6. Fellegi

Nous sommes d'accord avec de nombreux points défendus par Fellegi. Aux États-Unis, par exemple, les données sur les revenus dans les petites régions aident à déterminer la répartition des derniers publics. Ces données ont leurs propres faiblesses, que le redressement du recensement ne vient pas corriger. De même, on constate d'importants déplacements de population au cours de la période intercensitaire. L'amélioration des données sur le revenu ou la tenue d'un recensement au milieu de la décennie pourrait s'avérer plus utile que le redressement du recensement décennal.

Il existe néanmoins un point que nous aimerions soumettre à l'attention de Fellegi. Toute décision de redresser le recensement, que ce soit aux États-Unis ou au Canada, entraîne l'engagement d'importants frais d'organisation et encourage de ce fait le remplacement de la collecte de données par la modélisation.

En somme, les données réelles (comportant des lacunes réelles) seraient remplacées par des modèles mathématiques compliqués et mal testés. Ce n'est pas ce que nous considérons comme du progrès. (TRADE-UP) (Beran et coll. 1988)

7. Les séries du PEP

En 1980, il manquait beaucoup de données aux enquêtes qui ont servi à évaluer l'erreur du recensement. Chaque méthode retenue pour combler les données manquantes donne un taux différent de sous-dénombrement estimatif. Au bout du compte, le Bureau s'est retrouvé avec une douzaine de séries différentes de PEP, chacune d'elles proposant un taux de sous-dénombrement estimatif pour 66 secteurs géographiques (villes centrales, États sans leurs villes centrales, États dans leur ensemble). Une série s'identifie par une paire de chiffres, par ex. PEP 2-9 et PEP 10-8. Pour plus de détails et des arguments quant aux mérites des diverses séries, voir FN p. 4, EKT p. 929 et Fay et coll. (1988, p. 63).

8. Cressie

Cressie convient (p. 34) que "les données et les méthodes du recensement de 1980 étaient inadéquates pour redresser avec précision les chiffres pour l'ensemble du pays". Évidemment, bon nombre des arguments conviennent aussi à la décision de 1990, et l'opinion de Cressie à leur égard risque de se distinguer de la nôtre. Nous ne voulons pas amorcer ici une discussion détaillée des événements de 1990, mais nous pouvons donner une réponse sommaire à certains points soulevés par Cressie.

Analyse démographique

Cressie, à l'instar d'autres critiques, s'appuie sur des estimations tirées de l'analyse démographique, une technique qui fait appel aux dossiers administratifs (certificats de naissance, certificats de décès, etc.) pour obtenir une estimation indépendante de la population globale. Pour plus de détails, voir Fay et coll. (1988).

Que vaut l'analyse démographique? Aussi surprenant que cela puisse paraître, les organismes statistiques publics changent constamment d'idée sur le passé. Aussi le PNB estimatif pour

Rien n'est parfait, et il ne faut pas laisser le mieux être l'ennemi du bien

Fienberg affirme ce qui suit (p. 29):

Le cas des problèmes soulevés quand les hypothèses ne sont pas respectées constitue un thème familier dans divers écrits d'un des présents auteurs. Ici encore, les auteurs poursuivent ce thème relativement à l'équation linéaire employée pour effectuer le lissage. Ils semblent soutenir qu'il faut que toutes les hypothèses soient parfaitement justifiées, sinon "plus rien ne tient". Rien ne saurait être plus faux.

Hélas, notre position est plus compliquée que cela. Nous considérons le recensement comme imparfait, mais bon. Nous considérons les modèles de lissage comme fort douteux, et les arguments invoqués pour les défendre comme mauvais. Les tenants du redressement ont l'obligation d'annoncer leurs hypothèses et de produire des données susceptibles de les valider. Ils n'ont pas à établir des modèles à toute épreuve, mais il faut que soient étudiés les écarts par rapport aux hypothèses et leurs conséquences. Autrement, les algorithmes n'ont d'autre justification que la familiarité.

5. Le fardeau de la preuve

Comme en atteste l'échange avec Fienberg, les modélisateurs hésitent à accepter le fardeau de la preuve. Une fois posée une hypothèse, elle est considérée comme la réalité tant qu'elle n'a pas été infirmée. Une hypothèse infirmée peut même être considérée comme utile jusqu'à ce qu'elle soit remplacée par une autre hypothèse.

Les mots revêtent un sens particulier. Une hypothèse est "raisonnable" si les modélisateurs la considèrent comme telle. S'il y a remise en question, les modélisateurs se replient sur eux-mêmes. Cette introspection confirme la conclusion initiale; après tout, les hypothèses sont devenues partie intégrante de la documentation technique. Les modélisateurs s'indignent lorsque confrontés à ceux qui n'ont pas la foi. Si toutes les options "raisonnables" favorisent le redressement, les arguments contraires doivent forcément être "déraisonnables".

Si l'on en croit les textes publiés, les expériences intellectuelles des modélisateurs ne semblent pas particulièrement rigoureuses; de plus, l'argument en faveur du redressement emprunte parfois de curieux détours par rapport aux voies empiriques. Voici quelques illustrations. Un axiome du modèle de lissage d'Ericksen-Kadane est l'indépendance (voir les équations 1 à 6 de notre étude). Les calculs de la variance sont fonction de l'indépendance, car de faibles corrélations peuvent avoir une incidence cumulative énorme. Or, la variance détermine si le lissage constitue un outil précieux ou un obstacle. L'hypothèse de l'indépendance compte. D'après ce que nous pouvons constater, les tenants du redressement avancent surtout les arguments suivants en faveur de l'indépendance:

i) Les erreurs ne sont pas en corrélation parfaite. (Nous adaptons, au contexte présent, un argument proposé par Madansky en 1986, p. 29.)

ii) a) Le recensement de 1980 a été administré par plus de 400 bureaux de district, soit une moyenne de 8 par État. b) Autant que nous sachions, personne n'a donné à entendre qu'il y a effectivement eu tempête de neige en avril ou un autre événement qui a affecté le recensement dans les États avoisinants. c) Lorsque nous avons corrélé les estimations du PEP pour les villes avec les estimations correspondantes des États où elles étaient situées (par ex. Détroit avec le reste du Michigan), nous n'avons trouvé aucune trace de corrélation. [Étude EKT, p. 931; nous avons répondu au point b) en rappelant l'éruption du Mont St. Helens.]

iii) "Ils ne s'attendent certainement pas à ce qu'une personne quelconque accepte l'argument voulant que l'éruption du Mont St. Helens ait nu considérablement à la réalisation du recensement..." [Fienberg, p. 29]

4. Fienberg

Pour Fienberg, la question primordiale se définit comme suit:

L'exactitude des chiffres du recensement ainsi que le processus de redressement sont tous deux mis en cause. Et, c'est la différence importante du taux de sous-dénombrement, c.-à-d. la différence entre le sous-dénombrement des Noirs et le sous-dénombrement des personnes non noires et entre celui qui touche les personnes d'origine hispanique et le sous-dénombrement relatif aux personnes d'origine non hispanique qui est importante quand nous en venons à évaluer l'exactitude des chiffres du recensement. Cela est dû au fait que les chiffres du recensement sont généralement utilisés pour répartir les ressources entre des groupes de la population, ressources comme les sièges à la Chambre des représentants des Etats-Unis; les sièges dans les législatures des Etats; les fonds versés par l'administration fédérale et ainsi de suite. [p. 28, soulignement omis]

D'après nous, ce raisonnement est trompeur. L'argument porte effectivement sur les parts, plus précisément sur l'exactitude des parts calculées à partir de chiffres redressés et à partir du recensement. Les parts qui comptent sont toutefois celles qui concernent les secteurs géographiques: Etats, villes, comtés, etc. La part totale des Noirs ou des personnes d'origine hispanique dans la population américaine dans son ensemble compte beaucoup moins. Les sièges au Congrès sont répartis par Etat et, au sein de chaque Etat, par secteur géographique. Ils ne sont pas ventilés à l'échelle nationale parmi différents groupes raciaux ou ethniques. Par ailleurs, les recettes fiscales sont réparties parmi quelque 39,000 administrations locales, qui sont définies géographiquement, et non selon la race ou l'ethnie. La question fondamentale consiste à savoir si le redressement améliore l'exactitude des parts démographiques pour les secteurs géographiques, et non pour les groupes.

Fienberg est également trompeur à d'autres points de vue. En voici deux exemples.

i) Fienberg (p. 29). "[Freedman et Navidi] se concentrent sur la variation parmi l'ensemble complet de douze possibilités, dont je trouve certaines peu plausibles compte tenu des hypothèses sur lesquelles ils se basent." Cependant, nous avons effectivement étudié la variation parmi les séries d'estimations jugées supérieures par EKT, plutôt que parmi l'ensemble complet; voir les pages 7 et 8 et 14 et 15. Si nous avons procédé de la sorte, ce n'est pas parce que nous étions d'accord avec les choix d'EKT, mais parce que nous voulions rendre hors de propos le genre d'argument avancé par Fienberg. Cela ne l'a pas arrêté.

ii) Fienberg (p. 4). "J'ai lu le rapport de Ylvisaker (1991) qui a réexaminé les données provenant du recensement d'essai réalisé à Los Angeles en prévision du recensement de 1990, mais je n'ai pu y trouver la preuve qui, selon Freedman et Navidi, appuie leur prétention que le lissage augmente la variabilité." Ylvisaker a en effet procédé à une expérience d'auto-amorçage en utilisant des données de Los Angeles, où ont été menés un recensement d'essai et une enquête postcensitaire d'essai en 1986. Au niveau des secteurs de recensement, l'expérience a dégagé des erreurs-types d'échantillonnage plus importantes, en règle générale, pour les estimations lissées que pour les estimations brutes. (Voir le tableau 3 de Ylvisaker; le lissage a réduit l'erreur-type d'échantillonnage dans 19 secteurs de recensement sur 61, l'a augmentée dans 26 secteurs sur 61 et n'a eu aucun effet dans les 16 secteurs restants; au niveau des ilots, les effets vont dans le sens contraire, mais ils sont négligeables.) Pour l'ensemble de l'expérience, la comparaison s'établit comme suit (Ylvisaker, p. 7):

Erreur-type d'échantillonnage pour l'estimation lissée: 0.75.

Erreur-type d'échantillonnage pour l'estimation brute: 0.68.

Comme nous l'avons fait remarquer (p. 18), "le lissage peut accroître l'erreur d'échantillonnage".

RÉPONSE DES AUTEURS

1. Introduction

Après avoir formulé quelques généralités, nous répondons aux principaux points dégagés par chacun des critiques. Comme il existe un certain chevauchement, nous nous sommes efforcés de traiter chaque point une seule fois. À l'instar des autres participants, nous n'avons pas cessé d'apprendre au fil des ans – comme en témoigne d'ailleurs le présent échange de vues – mais nous n'avons pas modifié nos opinions quant aux questions fondamentales. Un fait demeure cependant incontestable: M. P. Singh, le rédacteur en chef, mérite des remerciements de toutes les parties.

2. Les grandes lignes du redressement

On a proposé de redresser le recensement au moyen de techniques de saisie-résaisie. Une personne est "saisie" lorsqu'elle est dénombrée dans le recensement; la "résaisie" a lieu à l'occasion d'une enquête spéciale par échantillonnage effectuée après le recensement. En 1980, cette enquête était désignée le "Post Enumeration Program" (PEP) (programme de contrôle postcensitaire). En 1990, on a parlé plutôt de "Post Enumeration Survey" (enquête post-censitaire, ou EP).

Ces enquêtes cherchent à préciser dans quelle mesure les personnes sont oubliées au moment du recensement ("oubliés bruts") ou dénombrées par erreur ("enregistrés erronés"). Parmi les enregistrés erronés, on trouve les bébés nés le lendemain du recensement et les personnes dénombrées à la mauvaise adresse, par exemple. Aux fins d'une première approximation, le sous-dénombrement net estimatif représente l'écart entre les oubliés bruts et les enregistrés erronés.

Un autre facteur vient compliquer la question. En 1980, l'erreur d'échantillonnage consistait, de l'avis de bien des observateurs, un problème assez grave pour empêcher l'utilisation directe des estimations de l'enquête. Aussi a-t-il fallu, selon les expressions de EKT, passer les "estimations d'échantillon" du PEP par un modèle de lissage afin d'obtenir des "estimations composées". En 1990, la terminologie avait changé: des "facteurs de redressement bruts" tirés de l'EP ont fait l'objet d'une modélisation dans le but d'obtenir des "facteurs de redressement lisés". Le problème de l'erreur d'échantillonnage est toutefois plus important qu'il y a dix ans. Pour plus de détails, voir Freedman (1991), U.S. Department of Commerce (1991a, p. 4.2-18) ou Wolter (1991).

3. Le recensement est mauvais, de sorte que les solutions de rechange doivent être meilleures

Bien des critiques ont proposé un argument qui, réduit à sa plus simple expression, revient à l'énoncé suivant: le recensement est mauvais; l'EP doit être meilleure; par conséquent, nous devrions procéder à un redressement. Il y a ici confusion: on considère le recensement et l'EP comme des solutions interchangeables. Pourtant, on ne peut pas choisir l'enquête plutôt que le recensement; au mieux, on peut tâcher d'utiliser l'EP pour corriger les lacunes du recensement. La question n'est donc pas de savoir si l'enquête est meilleure, mais plutôt de déterminer si elle convient aux fins auxquelles on la destine.

Le secrétaire du Commerce a formulé les enjeux comme suit:

Je concède les imperfections du recensement, mais la question fondamentale porte non sur le degré d'imperfection du recensement, mais sur l'aptitude de l'EP à les rectifier. Bien qu'il soit important de repérer les lacunes d'un recensement lorsqu'on planifie le prochain, il s'agit d'une démarche qui passe à côté de la question suivante: existe-t-il une preuve convaincante que le redressement est plus exact que le recensement? (TRADEACTION) (U.S. Department of Commerce, 1991a, p. 2.13)

indépendantes, et la compatibilité des estimations du sous-dénombrement avec les résultats de l'analyse démographique sont toutes des raisons qui justifient amplement un redressement. Freedman et Navidi placent les données redressées à un niveau plus élevé que les données non redressées. Ils tiennent pour acquis, malgré des dizaines d'années de compilation de renseignements au Census Bureau, que les données non redressées sont exactes et ils ne semblent pas être conscients de l'existence d'un biais important dans les régions. De plus, faute de données directes, ils supposent l'existence de biais appréciables dans les données du PEPF alors que les études du Census Bureau ne démontrent rien de ce genre (U.S. Bureau of the Census 1988, section 6F). Autrement dit, ils ne semblent pas placer les données non redressées et les données redressées sur un même pied d'égalité lorsqu'ils font leur analyse. Par conséquent, F et N sont capables d'"inventer" des problèmes et, en même temps, de négliger les vrais problèmes que cause le non-redressement des chiffres du recensement. Ils refusent le redressement sur cette simple base.

SOURCES ADDITIONNELLES

ERICKSEN, E.P. (1983). Affidavit présenté au U.S. District Court, Southern District of New York, dans *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

ERICKSEN, E.P. (1986). Commentaire sur Regression Models for Adjusting the 1980 Census, par D.A. Freedman et W.C. Navidi. *Statistical Science*, 1, 18-21.

ERICKSEN, E.P., et KADANE, J.B. (1986). Using administrative lists to estimate census omissions. *Journal of Official Statistics*, 2, 397-414.

ERICKSEN, E.P., et KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. Dans *Small Area Statistics: An International Symposium*. (Eds. R. Platek, J.N.K. Rao, C.E. Särndal et M.P. Singh). New York: John Wiley & Sons.

FREEDMAN, D.A. (1984). Témoignage donné au U.S. District Court, Southern District of New York, dans *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

KADANE, J.B. (1986). Commentaire sur Regression Models for Adjusting the 1980 Census, by D.A. Freedman et W.C. Navidi. *Statistical Science*, 1, 12-17.

TURKEY, J.W. (1983). Affidavit présenté au U.S. District Court, Southern District of New York, dans *Cuomo v. Baldrige*, 80 Civ. 4550 (JES).

U.S. BUREAU OF THE CENSUS (1985). The Coverage of Housing in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E1. Washington, DC: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1987). Programs to Improve Coverage in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E3. Washington, DC: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1988). The Coverage of Population in the 1980 Census, 1980 Census of Population and Housing Evaluation and Research Report PHC-80-E4. Washington, DC: U.S. Government Printing Office.

Si nous appliquons simplement l'équation (11) aux 66 régions sans faire de moyenne avec les estimations d'échantillon initiales, la variation des parts de population est la suivante: Groupe 1, + 0,36%; Groupe 2, + 0,20%; Groupe 3, - 0,56%. En utilisant l'équation (12), nous obtenons: Groupe 1, + 0,33%; Groupe 2, + 0,21%; Groupe 3, - 0,54%. Tandis que la différence entre les équations (11) et (12) s'explique facilement et est conforme à notre théorie de l'erreur dans le recensement, elle n'influe pas réellement sur les résultats finals.

Freedman et Navidi reviennent sur la question de savoir s'il est tout aussi bon d'utiliser la proportion de population urbaine que le taux de criminalité. Comme nous l'avons expliqué dans EKT, on obtenait un meilleur é.q.m. et des erreurs-types moindres avec le taux de criminalité. Dans leurs tableaux 7 et 8, F et N ne semblent pas arriver aux mêmes conclusions que nous. Cette divergence s'explique par le fait que F et N n'ont pas, comme nous le disions plus haut, reproduit correctement notre méthode de sélection. En utilisant des données non pondérées, ils n'ont pas reproduit notre méthode de régression, d'où le fait qu'ils ont obtenu des résultats différents. Comme leur méthode accorde plus d'importance aux villes, où les échantillons étaient plus petits et les estimations, par conséquent, moins précises, il n'est pas étonnant que la "proportion de population urbaine" occupe une place importante dans leur analyse. Freedman et Navidi pensent peut-être que notre décision de pondérer les observations par leur niveau de fiabilité estimé est une autre décision arbitraire. La pondération est une opération qui semble parfaitement correcte à nos yeux et qui est conforme aux principes directeurs que s'est donnés le Census Bureau pour 1990. Lorsque une observation semblait plus fiable, nous lui attribuions un poids plus grand. Toutefois, comme ils n'ont pas fait de pondération, F et N présentent une analyse qui est en grande partie différente de la nôtre et leurs résultats n'ont pas de rapport avec ce que nous avons fait. Les mêmes remarques valent pour les études de simulation qui ont été présentées dans l'article qui fait l'objet de ce commentaire et dans Freedman et Navidi (1986). Si F et N avaient pondéré les données, ils auraient probablement obtenu des résultats différents. Pourtant, le problème ne réside pas vraiment dans le fait que les variables choisies pour la régression diffèrent selon les auteurs. La vraie question est de savoir dans quelle mesure les estimations calculées à l'aide des différentes équations de régression varient entre elles. La réponse, comme nous l'avons vu plus haut, est que les taux de sous-dénombrement ne diffèrent pas sensiblement les uns des autres.

Conclusions

La principale observation d'EKT est probablement celle qui dit que pour n'importe quel ensemble de variables prédictives étroitement liées au sous-dénombrement, les estimations du sous-dénombrement varient peu selon les séries du PEP jugées raisonnables. En définitive, un changement de variable prédictive ou de série a relativement peu d'influence sur les estimations du sous-dénombrement. De même, nous n'accordons pas une grande importance aux résultats des simulations de F et N. Il n'est pas vraiment utile de savoir que des simulations différentes où l'on introduit des erreurs aléatoires donnent des "ensembles idéaux" de variables prédictives différents, à moins que la distribution des taux de sous-dénombrement s'en trouve modifiée, ce qui n'est pas le cas.

Faute de données directes d'évaluation, nous avons effectué une analyse de sensibilité pour étudier l'effet de diverses hypothèses sur les estimations. À notre avis, le remplacement d'une série du PEP par une autre aussi acceptable ou d'un prédicteur du sous-dénombrement par un autre tout aussi acceptable change peu de chose aux estimations. En outre, les résultats sont assez fidèles à ce qu'on pouvait prévoir compte tenu de la masse de renseignements dont nous disposons sur les problèmes de recensement. Dans les régions où le Census Bureau avait éprouvé plus de difficultés dans l'exécution du recensement, la proportion de personnes oubliées, la proportion d'enregistrements erronés et le niveau de sous-dénombrement étaient plus élevés. En définitive, nous pensons que l'existence d'une documentation abondante – et non contestée – sur les problèmes liés à l'exécution des recensements, de même que l'"invariabilité relative" des estimations par rapport au choix des variables dépendantes et

par régression et des estimations d'échantillon initiales [...] (TRADUCTION) (EKT, p. 935). Les observations ont été pondérées par l'inverse de l'erreur-type des estimations d'échantillon initiales. Cette pondération était indispensable car dans certains cas, comme la Caroline du Sud, les estimations d'échantillon étaient aberrantes et les variances élevées, et les échantillons pour les 16 villes étaient aussi plus petits, ce qui donnait des estimations moins précises. En pondérant ainsi les données, nous avons pu, par exemple, ramener de 24 à 12% la proportion de poids totaux attribués aux villes. Lorsque nous avons ajouté la variable de pauvreté à nos trois autres variables dans une régression pondérée, nous nous sommes rendu compte que son coefficient était moins que le double de l'erreur-type correspondante; c'est pourquoi nous avons exclu cette variable de l'analyse.

Freedman et Navidi confondent aussi une décision statistique avec des sentiments personnels. Au contraire, si la variable de pauvreté, avec son coefficient négatif, avait répondu à nos critères statistiques, elle aurait rehaussé nos estimations par de l'information utile et intéressante. En règle générale, il existe deux types de régions où le taux de pauvreté est élevé: les villes centrales qui comptent une forte proportion de minorités et les régions rurales d'Etats comme le Kentucky et la Virginie de l'Ouest où les minorités sont peu présentes. Les erreurs de recensement sont plus susceptibles de se produire dans ces régions que partout ailleurs, sauf que la nature de ces erreurs n'est pas la même dans les villes et dans les campagnes. Dans les villes, c'est la proportion des personnes oubliées qui est élevée, comme le montre le tableau 6 d'EKT, alors que dans les régions rurales, c'est la proportion d'enregistrements erronés qui est forte.

On peut se rendre compte de l'effet de l'inclusion de la variable de pauvreté en soustrayant l'équation (11) de F et N de l'équation (12), ce qui donne:

$$\text{différence dans l'ajustement de la série 2-9} = 2.23 + .041 \text{ min} - .010 \text{ crim} + .001 \text{ class} - .176 \text{ pauv.}$$

Dans les régions où le pourcentage de minorités et le pourcentage de la population vivant sous le seuil de la pauvreté sont tous deux élevés ou tous deux faibles, la différence n'est peut-être pas très notable. Dans les régions qui comptent une forte proportion de minorités mais où le taux de pauvreté est peut-être légèrement supérieur à la moyenne, le différence peut être positive, mais dans les régions qui comptent une faible proportion de minorités mais qui affichent un fort taux de pauvreté, la différence est négative. Pour ce qui a trait à l'ensemble des régions étudiées initialement, nous avons observé une différence supérieure à 1% dans quatre cas seulement sur 66 et une différence de 0.8 à 1% dans six cas seulement. Voici ces dix cas particuliers:

Région	Equation 12	Equation 11	Différence
Maryland R	2.3	1.2	1.1
Houston	5.3	4.2	1.1
Washington, DC	8.1	7.2	0.9
Cleveland	4.3	5.1	-0.8
Arkansas	-0.3	0.5	-0.8
Mississippi	1.0	1.8	-0.8
Dakota du Sud	0.4	1.3	-0.9
Kentucky	-1.3	-0.4	-0.9
Saint-Louis	5.5	6.6	-1.1
Boston	3.4	4.9	-1.5

et l'utilisation du modèle de régression semble produire une nette amélioration. Si l'on examine maintenant les 16 villes, on note des différences supérieures à 2% dans cinq cas. Dans ces cas précis, l'échantillon était plus petit et la moyenne pondérée se rapproche beaucoup plus de l'estimation par régression que de l'estimation d'échantillon originale. Même si F et N aimeraient mieux que nous ne calculions pas la moyenne pondérée, nous avons choisi de réserver un rôle, aussi modeste soit-il, aux données d'échantillon pour tenir compte des facteurs qui ne sont pas nécessairement inclus dans le modèle de régression. D'une manière ou d'une autre, même si nous persistons à dire que les hypothèses de notre modèle d'estimation sont raisonnables, nous pensons que l'argumentation devrait être axée plutôt sur la qualité des données du PEP.

Importe-t-il d'utiliser une série plutôt qu'une autre?

F et N répètent qu'il n'y a aucune raison de choisir une série du PEP plutôt qu'une autre. Au contraire, tandis qu'il peut être difficile de choisir une série parmi les huit jugées supérieures, nous avons de bonnes raisons de ne pas inclure la série 10-8 dans ce groupe. Ce n'est pas une solution d'exclure de l'analyse les personnes ayant déménagé – comme cela s'est fait dans le cas de la série 10-8 – sous prétexte que dans la CPS d'août, on a eu de la difficulté à déterminer l'adresse où vivaient en avril les personnes ayant déménagé. Comme le reconnaissent eux-mêmes Freedman et Navidi, et comme nous l'ont appris les résultats du PEP, le taux de sous-dénombrement est plus élevé chez les personnes ayant déménagé. Les faiblesses de la série 10-8 s'expriment aussi de deux autres manières. Premièrement, le taux de sous-dénombrement national calculé à l'aide de cette série (0.3%) est bien au-dessous du taux estimé au moyen de l'analyse démographique (1.4%). Deuxièmement, la variabilité entre les régions est déraisonnablement faible, comme on peut le voir dans le tableau 5 d'EKT. La série 10-8 engendre un transfert de parts de population comparable à celui obtenu avec la méthode synthétique B de Schirm et Preston, laquelle, même si elle représente une amélioration par rapport aux chiffres non redressés, est nettement insuffisante. C'est pourquoi la variabilité des estimations de la série 10-8 entre les 66 régions étudiées est trop faible. Par exemple, si nous attribuons le même poids à chacune des 66 régions, comme semblent l'avoir fait Freedman et Navidi, la variance inter-régions pour les estimations de la série 2-9 est plus de deux fois supérieure à la variance inter-régions correspondante pour les estimations de la série 10-8. Par rapport à la moyenne nationale, les estimations de la série 10-8 sont trop faibles dans les régions à fort taux de sous-dénombrement et trop élevées dans les régions où le taux de sous-dénombrement est faible. Il est donc peu étonnant que les résidus de la régression soient un peu plus petits pour la série 10-8 que pour la série 2-9, et le tableau 7 de F et N n'a pas de signification réelle.

Quelles variables explicatives utiliser?

Freedman et Navidi croient que lorsque la série 2-9 servait de variable dépendante dans la régression, nous n'avons pas respecté nos propres règles pour le choix des variables indépendantes. Ils prétendent que nous aurions dû ajouter le "pourcentage de la population qui vit sous le seuil de la pauvreté" aux trois variables que nous avons choisies – à savoir le pourcentage de minorités, le taux de criminalité et la proportion de la population recensée selon la méthode classique – parce que les coefficients des quatre prédicteurs de "leur" équation (11) étaient plus de deux fois supérieurs à l'erreur-type correspondante et que cette équation avait un é.q.m. relativement moins élevé. Ils affirment en plus que parce que le coefficient de la variable pauvre était négatif, nous avons rejeté l'équation que nous aurions dû normalement choisir suivant nos critères statistiques. Autrement dit, ils prétendent que nous avons laissé prévaloir nos préconceptions au détriment de notre sens statistique.

L'inconvénient de la critique de F et N est qu'ils n'ont pas reproduit correctement notre méthode de sélection. Comme nous l'avons expliqué dans EKT et dans d'autres ouvrages (Ericksen et Kadane 1985; 1987, section 6), "[...] L'estimation du taux de sous-dénombrement que nous avons calculée est, sous une forme matricielle, la moyenne pondérée d'une estimation

d'améliorer les chiffres du recensement pour de grandes régions, on pouvait espérer faire de même en règle générale pour des sous-régions. Depuis lors, des progrès ont été faits tant sur le plan théorique qu'empirique (Ericksen et coll. 1991, annexe H; Wolter et Causey 1991).

Mise en moyenne et analyse de sensibilité

Freedman et Navidi affirment que "[...] L'important dans une série du PEP n'est pas la moyenne mais l'écart, car c'est l'écart [...] qui illustre l'effet d'hypothèses de modélisation de différences sur une même série de données [...] (p. 12). Nous contestons l'affirmation de F et N à deux points de vue. Premièrement, nous croyons que la moyenne et l'écart sont **aussi importants** l'un que l'autre, et nous avons fait une analyse des deux. Deuxièmement, et plus important encore, nous avons utilisé une mesure différente pour l'écart, soit l'écart quadratique moyen (é.q.m.) au lieu de l'intervalle de variation. F et N justifient peu le choix de leur mesure. Quant à nous, nous préférons l'é.q.m. parce qu'il tient compte de toutes les données et que le "carté de l'erreur" permet de donner plus de poids aux grosses erreurs. Nous avons noté que "[...] L'écart quadratique moyen pour les 792 résidus est de 0.59. Par contraste, l'écart quadratique moyen des 66 "effets de région" est de 1.60. L'effet de région est plus que le double de l'écart quadratique moyen dans 47 cas sur 66 [...] (EKT, p. 938). Nous avons aussi montré que lorsque nous attirons notre attention aux huit séries jugées supérieures, l'écart quadratique moyen était de 0.33 et l'effet de région était plus de deux fois supérieur à l'é.q.m. dans 59 cas sur 66.

Nous pensons que l'utilisation de l'intervalle de variation dans le tableau 6 et la figure 1 de F et N est injustifiée pour une autre raison. Même parmi les séries d'avril jugées supérieures, on note des différences dans le taux de sous-dénombrement national. Si, comme le disent F et N, nous nous préoccupons des transferts de parts de population, nous devrions aussi être attentifs aux écarts par rapport à la moyenne nationale, comme dans "nos" tableaux 10 et 11. Par exemple, l'estimation du sous-dénombrement pour la Floride est 2.63% selon la série 2-20 et 1.42% selon la série 3-8, soit un écart de 1.21 point. Si nous soustrayons de ces taux les taux nationaux respectifs de 1.9 et de 1.0%, nous obtenons des estimations de 0.73 et 0.42%, soit un écart de 0.31% seulement. L'utilisation de cette statistique amoindrit la corrélation mise en évidence par Freedman et Navidi.

Hypothèses

Freedman et Navidi prétendent que certaines des hypothèses qui sous-tendent notre modèle de régression sont "non vérifiées" et "peu plausibles". Comme nous l'avons déjà soutenu, que ce soit dans EKT ou dans d'autres ouvrages (Ericksen 1986; Kadane 1986) nous croyons que nos hypothèses sont réalistes et qu'elles reposent sur un bloc de connaissances accumulées depuis des décennies. F et N prétendent que notre modèle donne de meilleurs résultats que la méthode synthétique à la seule condition qu'il utilise l'information additionnelle d'une manière sensée, ce qui nous ramène aux hypothèses. En dépit de ces affirmations, nous estimons que nos hypothèses sont sûrement raisonnables et certainement plus réalistes que les hypothèses sur lesquelles repose la décision de ne pas faire de redressement.

Par ailleurs, nous pensons qu'il est possible de surestimer le rôle de la modélisation dans l'estimation du sous-dénombrement. Dans le cas des petites régions, il faut une certaine forme de modélisation. Cependant, en ce qui concerne les 66 régions qui faisaient l'objet de notre article, la modélisation n'a pas vraiment changé grand-chose. Par exemple, si nous comparons les résidus moyens de "notre" tableau 1, qui sont la moyenne des résidus des estimations tirées des huit séries jugées supérieures, au résidu moyen correspondant des huit estimations d'échantillon, nous constatons ce qui suit en ce qui a trait aux 50 États. Quarante-six résidus sont à moins d'un point de pourcentage les uns des autres tandis que quarante-huit sont à moins de 1.5 point les uns des autres. Les deux États restants, soit la Caroline du Sud et le Tennessee, semblent avoir des estimations d'échantillon erronées, comme nous l'expliquions dans EKT,

du PEP mais avec ceux d'une étude indépendante du Census Bureau qui portait sur la ville de New York et qui a révélé l'existence d'une corrélation forte et négative entre les taux de sous-dénombrement et les taux de réponse par la poste dans tous les districts de recensement (Ericksen et Kadane 1986).

Freedman et Navidi fondent leur argument sur les différences observées entre les distributions corrigées qui découlent des diverses séries du PEP. Nous ne croyons pas que ces différences sont pertinentes car nous savons que les huit "séries du PEP jugées supérieures", comme les estimations synthétiques A, plus raisonnables, différeront non seulement des estimations synthétiques B mais aussi des quatre séries du PEP jugées moins intéressantes. Si l'on prend les six séries "jugées supérieures" qui reposent sur les données d'avril, la valeur moyenne de l'é.q.m. est de 0.07%. L'écart entre ces séries et les deux séries "jugées supérieures" qui reposent sur les données d'août est plus grand que cela, mais nous avons expliqué dans notre article pourquoi nous pensions que les données d'avril et celles d'août étaient différentes. Il y a un point plus important encore toutefois: les quatorze formes de redressement ont pour effet d'améliorer les chiffres du recensement en transférant des parts de population des régions où il n'y a pas eu beaucoup de problèmes de recensement vers celles où les problèmes étaient nombreux. Le fait que certaines des méthodes de redressement (notamment la méthode synthétique B) modifient relativement peu les parts de population n'est pas une raison pour s'opposer à tout redressement.

F et N formulent d'autres objections, auxquelles nous pouvons répondre facilement. Premièrement, Freedman et Navidi ne semblent pas appuyer l'idée d'intégrer des données de plusieurs sources. Pourtant, il n'y a rien qui laisse croire que la combinaison de sources d'information soit une opération incohérente ou inhabituelle. Comme nous partons du principe que de l'information additionnelle est généralement une bonne chose, nous ne voyons pas comment F et N pourraient avoir raison sur ce point. Deuxièmement, constatant que la méthode démographique ne permet pas une décomposition géographique détaillée du sous-dénombrement, F et N la mettent de côté comme si elle était parfaitement inutile. À nos yeux, la méthode démographique fournit au moins deux renseignements importants. Elle produit une estimation juste du taux de sous-dénombrement national et elle révèle une covariable puissante: le sous-dénombrement est plus marqué chez les Noirs que chez les Blancs. Il ne faut pas croire que ces estimations sont sans erreur, mais elles nous donnent de bonnes raisons de penser que chacune des séries du PEP jugées supérieures concorde avec les observations obtenues à l'aide de la méthode démographique. Enfin, F et N trouvent superflu "notre" tableau 6, où nous montrons que le rapport entre la proportion des personnes oubliées et la proportion des enregistrements erronés est élevé dans les régions qui ont un fort taux de sous-dénombrement. En ce qui concerne le tableau 5 d'EKT, nous trouvons qu'il est conforme non seulement au tableau 6 mais aussi aux résultats de l'analyse démographique, ce qui nous amène à croire davantage en l'utilité du PEP. On appelle cet argument la "validité convergente" et on le retrouve souvent en sciences sociales. Notons aussi que les séries que nous jugeons moins intéressantes en raison du peu de vraisemblance de leurs hypothèses – séries 10-8, 14-8, 14-9 et 14-20 – concordent moins bien avec les résultats de l'analyse démographique. Voici nos constatations:

1. Validation des huit séries jugées supérieures, du fait que le taux de sous-dénombrement national correspondant et les résultats relatifs aux régions à forte concentration de Noirs sont conformes aux résultats de l'analyse démographique;
2. Données permettant d'affirmer que la série 10-8, la série-repère de Freedman et Navidi, est effectivement une série "aberrante".

Pouvons-nous espérer améliorer les chiffres du recensement dans les petites régions?

Dans notre témoignage devant le tribunal, nous cherchions surtout à montrer qu'il était possible d'améliorer les chiffres du recensement pour les 66 régions définies comme les secteurs-échantillon du PEP. Dans un autre document, Tukey (1983) a montré que s'il était possible

Par ailleurs, Freedman et Navidi ne peuvent construire leur propre argumentation sans se fonder sur des assertions qui sont non vérifiées ou qui reposent sur les données mêmes du PEP que F et N nous reprochent d'utiliser. Voici quelques exemples de ces assertions:

1. On relève toujours un certain niveau de sous-dénombrement, modeste, dans les recensements (p. 3).
2. Les fiches du recensement comprennent aussi un faible pourcentage d'enregistrements erronés (p. 4).
3. Les taux de sous-dénombrement établis à l'aide des estimations du PEP risquent d'être entachés d'un biais par excès (p. 13).
4. L'éruption du mont St. Helens a engendré une corrélation des erreurs entre le recensement et le PEP (p. 13).
5. Les données manquantes ont introduit un biais dans les données du PEP (p. 14).
6. Les membres de minorités qui vivent dans les villes centrales ont des chances d'être différents de ceux qui vivent dans les banlieues (p. 18).
7. Le taux de sous-dénombrement était relativement élevé dans les régions ayant fait l'objet d'un recensement classique (p. 19).

F et N semblent croire qu'à défaut d'une information directe abondante sur la qualité des données du PEP, il faut s'abstenir de redresser les chiffres du recensement étant donné que des hypothèses différentes aboutissent parfois à des résultats différents. Or, cet argument fait abstraction de la riche documentation qui existe sur les erreurs dans le recensement. Nous avons traité abondamment, dans l'article désigné par les lettres EKT comme dans d'autre ouvrages (Ericksen 1983; Ericksen et Kadane 1985), les problèmes que l'on rencontre dans le recensement et d'autres (Citro et Cohen 1985; U.S. Bureau of the Census, 1985, 1986 et 1988) ont obtenu les mêmes résultats que nous. À notre avis, il ne fait aucun doute que l'information additionnelle tirée des données du PEP aurait pu servir à redresser – et à améliorer du même coup – les chiffres du recensement de 1980 si l'on considère la documentation abondante sur les problèmes liés à l'exécution des recensements, la relation géographique qui existe entre ces problèmes et les taux de sous-dénombrement élevés (selon le PEP) et la compatibilité des séries du PEP entre elles et leur compatibilité avec les résultats de l'analyse démographique. Cela résume notre point de vue général. Dans les sections qui suivent, nous examinons quelques-uns des arguments présentés par Freedman et Navidi.

Les méthodes simples de redressement améliorent-elles les chiffres de recensement?

Dans la section 2 de leur article, F et N font une critique de 'notre' tableau 5, qui visait à montrer la concordance de quatorze méthodes de redressement différentes. Chacune de ces méthodes a pour effet de transférer des parts de population des régions habitées surtout par des Blancs à l'extérieur des villes, où le recensement ne posait pas vraiment de problèmes, vers les grandes villes centrales qui comptent une forte proportion de minorités, où le recensement posait beaucoup de problèmes. F et N concluent: '[Le tableau] n'indique pas si l'une ou l'autre de ces méthodes accroît réellement l'exactitude des chiffres du recensement et avec raison, car il n'existe aucun critère externe permettant de le vérifier.' (p. 6). Si F et N entendent par cela que la population 'vraie' est inconnaisable, alors leur argument est sûrement exagéré et aucune méthode de redressement ne pourra jamais répondre à leurs exigences.

Dans EKT, nous avons eu recours à Schirm et Preston (1987) pour montrer que l'on pouvait améliorer les chiffres du recensement à l'aide d'une méthode synthétique simple (en l'occurrence, la méthode synthétique B). Comme ces auteurs ont écrit eux aussi un commentaire sur l'article de Freedman et Navidi, nous ne reprendrons pas ici leurs arguments. Étant donné l'amélioration observée avec la méthode synthétique B, nous devrions nous attendre à une amélioration encore plus grande avec des hypothèses plus réalistes comme celle voulant que le dénombrement des minorités se fasse plus difficilement dans les régions où les problèmes de recensement sont plus nombreux. Ces hypothèses concordent non seulement avec les résultats

COMMENTAIRES

EUGENE P. ERICKSEN et JOSEPH B. KADANE¹

Juge Sprizzo: "Si je comprends bien, l'erreur-type doit être un nombre fixe que vous soustrayez de vos résultats et le reste est supposé essentiellement mesurer le degré d'exactitude de ce que vous calculez?"

Le témoin (David Freedman): "Désolé, monsieur le juge, mais ce que vous dites n'est pas tout à fait juste." (Cuomo v. Baldrige: 2629). (TRADUCTION)

Nous sommes heureux de pouvoir poursuivre le débat engagé avec messieurs Freedman et Navidi (F et N). Bien que la décision du juge Sprizzo remonte à plus de quatre ans déjà, les questions statistiques qui ont été au coeur du litige sont des questions fondamentales qui méritent toujours notre attention. Et ceci est d'autant plus vrai que les jugements scientifiques finals appartiennent aux statisticiens et aux démographes plutôt qu'aux juges et aux politiciens. Dans leur article, Freedman et Navidi exposent de nouveau leur position sur la question du redressement, présentent de nouveaux arguments et tentent de recourir à la décision du juge Sprizzo pour appuyer leur position scientifique. De notre côté, nous allons réexaminer certains points fondamentaux en prenant soin de répéter et de clarifier au besoin notre position dans le but de montrer qu'un redressement des chiffres du recensement de 1980 aurait donné des résultats plus exacts.

Les points de désaccord qui existent entre Freedman et Navidi et nous sont fondamentaux et il est juste de dire que ces points de désaccord touchent les fondements de l'inférence statistique. Dans leur conclusion, F et N écrivent: "[. . .] l'efficacité de l'un ou l'autre des redressements proposés par EKT repose sur des hypothèses non vérifiées et peu plausibles [. . .]" (TRADUCTION) (p. 20). Au contraire, nous croyons que nos hypothèses sont réalistes et qu'elles sont vérifiées par des dizaines d'années d'expérience en recensements, comme nous le verrons plus loin. Pour leur part, F et N ne font guère plus que s'inquiéter de ce que certaines hypothèses pourraient ne pas être vraies. Or, il faut plus que cela pour critiquer un argument statistique. En règle générale, une hypothèse n'est jamais parfaitement exacte; ce qui compte, c'est de savoir dans quelle mesure elle est inexacte et en quoi cet écart peut influer sur l'usage que l'on veut faire des données. Pour réfuter une hypothèse, il faut à tout le moins montrer qu'il y a d'autres hypothèses, jugées tout aussi réalistes sinon plus réalistes, qui mènent à des conclusions sensiblement différentes. F et N ne font rien de cela. En outre, bien qu'ils s'arrêtent aux différences mineures qui existent entre les diverses possibilités de redressement, ils ne tentent aucunement de démontrer que les chiffres redressés sont moins exacts que les chiffres non redressés.

Une partie importante du différend porte sur la pertinence d'utiliser les données dont nous disposons au sujet du recensement. F et N n'accordent aucune importance au fait que le recensement s'effectue plus difficilement dans certaines régions et ne prennent pas en considération non plus le fait que les taux de sous-dénombrement calculés à l'aide des données du PEP sont plus élevés dans les régions où le taux de réponse par la poste est relativement faible et la proportion de données manquantes plus élevée et où il est plus difficile de respecter le taux d'échantillonnage spécifié pour les questionnaires complets du recensement. F et N n'accordent pas plus d'importance aux résultats des analyses démographiques effectuées pour chaque recensement depuis 1940, selon lesquels les différences de taux de sous-dénombrement entre les origines raciales sont constantes. Cette information n'est pas pertinente aux yeux de F et N, et ceux-ci s'empressent de nous critiquer dès que nous fondons sur des hypothèses "non vérifiées", quelque réalistes ou justifiées qu'elles soient. Ils ne nous disent pas non plus ce qu'ils entendent par "vérification".

¹ Eugene P. Ericksen, Temple University, Philadelphia, PA USA, and Joseph B. Kadane, Carnegie-Mellon University, Pittsburgh, PA USA 15213.

relatives à un modèle tout à fait différent. À mon avis, la modification des hypothèses influe sur les résultats, et si nous n'avons aucun moyen de déterminer quelle série de résultats se rapproche le plus de la réalité, vaut mieux ne pas redresser. C'est le message que visent à nous communiquer Freedman et Navidi et que j'appuie sans réserve.

SOURCE ADDITIONNELLE

HENGARTNER, N., et SPEED, T. P. (1992). Assessing between-block heterogeneity within poststrata of the 1990 Post-Enumeration Survey. Soumis pour publication dans le *Journal of the American Statistical Association*.

COMMENTAIRES

T.P. SPEED¹

Freedman et Navidi se demandent s'il aurait fallu redresser les chiffres du recensement de 1980 et concluent que non. J'interprète leur réponse comme une manière de dire qu'ils ne sont pas convaincus de l'efficacité des méthodes de redressement qui leur ont été présentées jusqu'à maintenant et non comme un refus d'envisager le problème du sous-dénombrement ni comme une question de redressement il y a environ quatre ans, peu après mon arrivée aux États-Unis. J'ai lu les articles qu'ont rédigés les principaux intéressés et j'ai eu l'occasion, récemment, d'analyser des données d'îlots recueillies dans le cadre de l'Enquête postcensitaire de 1990. J'en arrive à la même conclusion que Freedman et Navidi: il n'existe aucune donnée permettant d'affirmer qu'un redressement donnera des résultats satisfaisants au niveau proposé.

Il y a deux choses qui me frappent particulièrement lorsque j'examine les arguments de ceux qui préconisent un redressement. Premièrement, les adeptes du redressement n'ont jamais recours à des "données de terrain" pour démontrer hors de tout doute que les redressements améliorent vraiment les chiffres du recensement et deuxièmement, ils ne se servent pas des données disponibles pour vérifier les principales hypothèses sur lesquelles reposent les méthodes de redressement.

En 1990, le redressement devait se faire au niveau de l'îlot de recensement. Cela signifie que l'on aurait redressé l'effectif de chacun des 6,5 millions d'îlots en se fondant sur les données d'un échantillon d'environ 12,000 îlots. Le redressement de l'effectif recensé d'un îlot consiste à ajouter ou à soustraire des personnes qui ont des caractéristiques particulières avant de passer aux étapes d'agrégation suivantes. Cela se serait fait à l'aide de méthodes qui reposent sur des hypothèses non vérifiées et peu vraisemblables concernant le mécanisme de sous-dénombrement. La plus notable de ces hypothèses est celle voulant que les taux de sous-dénombrement sont constants pour chacun des 1,392 sous-groupes démographiques appelés "strates *a posteriori*" et définis selon la région, l'origine raciale, le sexe, l'âge et le mode d'occupation. Un de ces groupes comprend, par exemple, tous les locataires de sexe masculin et d'origine hispanique non noir âgés de 30 à 44 ans et vivant à Los Angeles ou dans des villes centrales de la division du Pacifique (Californie, Oregon, Washington, Alaska, Hawaï). Un autre comprend toutes les femmes propriétaires âgées de 20 à 29 ans et vivant dans les villes centrales de 250,000 habitants et plus (sauf New York City) de la division de l'Atlantique centre (New York, New Jersey, Pennsylvanie) et qui ne sont ni noires, ni d'origine hispanique, ni asiatiques, ni originaires des îles du Pacifique. L'analogie avec les modèles de régression présentés dans l'article est évidente.

Une analyse de données d'îlots recueillies à Détroit et au Texas dans le cadre de l'Enquête postcensitaire de 1990 a montré que l'hypothèse des taux de sous-dénombrement constants à l'intérieur des 1,392 strates *a posteriori* ne se vérifie guère plus qu'une autre hypothèse bien différente, à savoir que le sous-dénombrement varie selon les îlots mais qu'il est le même pour toutes les strates d'un îlot. Ce modèle dual aurait donné des résultats différents pour ce qui regarde le redressement au niveau de l'îlot. Il est toutefois difficile de faire une analyse car les effectifs qui correspondent aux diverses combinaisons "îlot-strate" sont minimes, sinon nuls, et hétérogènes. Pour plus de détails, se référer à Hengartner et Speed (1992). Évidemment, j'ignore lequel des deux modèles de sous-dénombrement – variation selon les strates ou selon les îlots – est le meilleur; il nous faudrait des "données de terrain" pour le savoir. Toutefois, compte tenu des données disponibles, nous constatons que certaines hypothèses clés concernant le modèle de sous-dénombrement de 1990 ne se vérifient guère plus que celles

¹ T.P. Speed, Département de statistique, Université de Californie, Berkeley, CA U.S.A. 94720.

vue politique, puisque les parties intéressées désiraient affecter les sièges et les ressources disponibles selon les chiffres qui leur seraient les plus favorables. On courrait le danger que les professionnels du recensement, qui disposeraient de plusieurs estimations, seraient soumis à une pression politique afin de choisir des estimations favorables à l'un ou l'autre groupe. Toutefois, si nous voulons estimer des populations avec précision, c'est un bon principe que de baser les estimations sur plusieurs enquêtes qui s'appuient mutuellement plutôt que sur une seule. Le danger de se fier à une seule liste l'emporte sur les dangers et les difficultés que comporte le fait de combiner des renseignements provenant de sources différentes. En particulier, la seule façon de déterminer l'exacritude d'une enquête consiste à en comparer les résultats à ceux d'une autre enquête sous une forme ou une autre. Il faudrait remarquer que même pour le recensement dont les chiffres 'ne sont pas redressés', les estimations de la population ne sont pas le fruit d'un simple comptage à partir d'une liste de dénombremment. Divers genres de procédures d'imputation utilisées pour traiter le cas des données manquantes introduisent des personnes fictives dans le total.

Il se peut que Freedman et Navidi aient raison quand ils affirment que les chiffres obtenus dans le cadre du PEP de 1980 étaient trop peu fiables pour qu'on puisse les utiliser afin de redresser les chiffres du recensement. Il se peut qu'ils aient eu raison quand ils ont dit qu'un redressement synthétique à l'échelle nationale produirait des résultats trop grossiers et qu'on n'a pas démontré qu'un tel redressement améliorerait l'exacritude. Cependant, les dizaines de millions de personnes oubliées et d'enregistrements erronés laissent supposer qu'un certain genre de redressement des chiffres pourrait améliorer l'exacritude. Il y a beaucoup matière à amélioration. Nous pourrions nous tromper en estimant l'existence ou le lieu de résidence de quelques millions de personnes et disposer encore de chiffres comparables aux résultats bruts du dénombremment.

J'ai quelques questions pour Freedman et Navidi.

- (1) Sont-ils d'accord avec ces estimations de 13 millions de personnes oubliées et de 17 millions d'enregistrements erronés?
- (2) La différence du taux de sous-dénombremment entre les Noirs et les Blancs est-elle de 4% à l'échelle nationale?
- (3) L'EP est-elle un outil utile pour évaluer l'exacritude du recensement original?
- (4) L'échantillon de suivi de l'EP devrait-il être utilisé pour corriger le recensement original, non seulement dans les cas précis de personnes dénombrées par erreur et de personnes omises découvertes quand on compare les enquêtes, mais aussi en projetant, à tout le recensement, les différences dans le taux de sous-dénombremment découvertes dans l'échantillon utilisé pour le suivi? Dans l'affirmative, comment?
- (5) Si l'EP ne donne pas des résultats satisfaisants, comment l'enquête de suivi devrait-elle être conçue afin qu'on puisse en utiliser les résultats pour redresser les chiffres du recensement?

de scepticisme et de prudence avant de croire cette troisième vérité. En particulier, les estimations de la population véritable¹ faites par le Bureau sont toutes des variations sur les estimations selon l'EP, dans lesquelles on accepte l'exacitude et la faisabilité de base de l'EP, il est donc fort peu probable qu'elles ne concordent pas avec l'EP comparativement à ce qui se produit dans le cas du dénombrement. Seule une méthode qui n'est pas aussi étroitement liée à l'EP peut permettre de trouver que ses résultats sont inférieurs à ceux du recensement. Les résultats de l'analyse démographique sont loin de concorder entièrement avec ceux de l'EP et cette méthode ne fournit que des renseignements à l'échelle nationale mais, dans l'ensemble, l'analyse démographique appuie l'EP plutôt que le dénombrement.

J'ai effectué quelques analyses de sensibilité pour évaluer dans quelle mesure les estimations et les marges d'erreur selon l'EP devraient être erronées pour que les résultats du recensement leur semblent comparables. Les calculs font appel aux estimations, selon l'EP, du sous-dénombrement au niveau de l'Etat avec divers multiplicateurs, aux marges d'erreur selon l'EP avec divers multiplicateurs et ils supposent que les chiffres véritables pour les Etats sont tirés de distributions normales avec les sous-dénombrements et les marges d'erreur selon l'EP multipliés pour les Etats. Je considère que la réalité pour les différents Etats est indépendante, ce qui est certainement loin d'être exact. Toutefois, l'indépendance n'aura pas un effet considérable sur les proportions pour chaque Etat, de sorte que cela ne modifiera pas beaucoup l'erreur de répartition moyenne selon le recensement et selon l'EP, la variabilité de l'écart sera sous-estimée.

Les résultats présentés dans le tableau 4 montrent que les estimations selon l'EP doivent être entachées d'une erreur considérable avant que les résultats du recensement commencent à leur être comparables. Si l'on accepte les taux et les marges d'erreur selon l'EP, le recensement entraîne une erreur de répartition de 4 sièges, l'EP d'un siège. Si l'on divise par deux les taux de surdénombrement selon l'EP, en conservant les marges d'erreur fixes, le taux d'erreur de répartition selon le recensement est de 2,5 sièges et de 1,5 selon l'EP, et le recensement donnera de meilleurs résultats dans environ 40% des cas véritables.

Toutefois, cette analyse s'apparente aux analyses des fonctions de perte en ce sens qu'elle utilise l'EP comme point de départ.

Je me demande s'il ne serait pas utile de faire une distinction entre le *dénombrement*, au cours duquel on dresse la liste des noms et des adresses, les *chiffres du recensement*, obtenus en comptant le nombre de personnes dans diverses localités selon la liste et les *estimations* obtenues à l'aide de méthodes statistiques basées sur diverses sources de renseignements comme la démographie et des listes supplémentaires. Cela semblerait peut-être la discorde du point de

Tableau 4

Erreur de répartition attribuable au dénombrement et au redressement à l'aide de l'EP, quand les chiffres véritables concordent avec les taux de sous-dénombrement et avec les marges d'erreur selon l'EP, avec divers multiplicateurs pour les sous-dénombrements et les marges d'erreur.

(Calcul basé sur 100 totaux véritables simulés.)

Multiplicateur pour le sous-dénombrement selon l'EP dans chaque Etat	Multiplicateur pour la marge d'erreur dans chaque Etat	Sièges mal répartis avec les chiffres du recensement	Sièges mal répartis avec les chiffres redressés	Ecart-type de la différence
1	1	3,8	1,1	1,2
0,5	2	3,7	2,0	1,3
0,5	0,5	2,8	1,2	1,3
0,5	1,0	2,5	1,6	1,4
0,75	0,75	3,3	0,9	1,3
0,75	1,50	3,3	1,6	1,3

Il ne suggère que la fonction de perte appropriée pour évaluer l'exacritude de la répartition des sièges n'est pas le carré de l'erreur, qui est un élément comme de point de vue statistique pour combiner les variances et les carrés des biais, ni les estimations du nombre d'États ou de localités pour lesquels les chiffres redressés du recensement donnent un meilleur résultat. Pour la répartition des sièges, la fonction de perte devrait être la somme de la valeur absolue des différences entre les proportions estimées et les proportions véritables dans les différents États, parce que cette somme représente le nombre de personnes qui ont effectivement été mal réparties par les estimations et elle correspond, au niveau de l'État, au nombre de sièges mal réparties.

Bien que les proportions pour les États présentent un intérêt primordial, étudions tout d'abord la population des États. Si le taux de sous-dénombrement véritable dans un État est de 2%, alors le recensement donne de meilleurs résultats que l'estimation seulement quand le taux de sous-dénombrement estimé est inférieur à 0% ou supérieur à 4%. Cela se produit, avec une probabilité de 50%, quand l'erreur-type de l'estimation est d'environ 3%; ainsi, même une estimation très imprécise du sous-dénombrement suffit pour que l'estimation redressée donne un résultat légèrement supérieur. Le recensement possède le même écart prévu par rapport à la réalité que l'estimation quand l'erreur-type est 2.2 et le même carré de la différence prévue quand l'erreur-type est 2. Ces résultats nous permettent de formuler la règle simple suivante: si le taux véritable est de 2%, on obtient de meilleurs résultats que le recensement si l'on peut estimer le taux véritable avec une erreur-type de 2%. Quand on estime des proportions de populations, plutôt que des populations, les calculs pertinents portent sur les différences du taux de sous-dénombrement pour les divers États, la différence entre le sous-dénombrement pour l'État et le sous-dénombrement à l'échelle du pays, (pas la différence entre les races); ainsi, les chiffres redressés sont meilleurs que les chiffres du recensement dans les États où la vraie différence entre le taux de sous-dénombrement pour l'État et le taux de sous-dénombrement pour le pays dépasse l'erreur-type de la différence.

Selon cette règle, et en acceptant les estimations de 1990 du Bureau of the Census pour les taux de dénombrement et les marges d'erreur basées sur l'enquête postcensitaire, on estime que, dans 24 États sur 50, le dénombrement donnera de meilleurs résultats pour ce qui est d'estimer les proportions. Il faut toutefois remarquer que la perte globale estimée est pas mal meilleure dans le cas des chiffres redressés que pour ceux du recensement, parce que les données redressées donnent une meilleure estimation pour les États avec des différences importantes (positives ou négatives) dans le taux de sous-dénombrement; quand le recensement donne de meilleurs résultats, c'est à peine le cas; quand le redressement donne de meilleurs résultats, il arrive souvent que ces derniers soient de beaucoup supérieurs. Le fait que pour 24 États sur 50 on estime que le redressement n'améliore pas les chiffres du recensement ne devrait donc pas causer trop d'émotion; cela signifie tout simplement que, pour beaucoup d'États, le taux de sous-dénombrement estimé est très près de la moyenne nationale et qu'il n'y a pas d'avantage à redresser les chiffres et que cela n'entraîne pas, non plus, un changement considérable. Nous continuons d'estimer que les chiffres redressés permettent de répartir correctement un nombre plus important de personnes que ce n'est le cas pour le dénombrement. Le tableau 4 présente certaines estimations des erreurs quand les chiffres du recensement basés sur l'EP sont inexacts de diverses façons.

Le Bureau a produit un certain nombre de sous-dénombrements estimés, avec des marges d'erreur, dans les divers États. J'utilise 'la méthode choisie pour l'EP' (que je désignerai dorénavant par l'abréviation EP) dans le rapport du Undercount Steering Committee (le comité directeur du sous-dénombrement) du 21 juin 1991. Nous possédons maintenant deux vérités, le dénombrement et les chiffres de l'EP. Laquelle est exacte? Bien, il nous faut une troisième vérité, une qui ne sera pas contestée pour décider.

Nous ne possédons pas cette troisième vérité. Les diverses évaluations de suivi de l'EP, le modèle de l'erreur totale, les analyses de fonctions de perte, les analyses de robustesse constituent toutes des tentatives pour trouver quelle serait cette troisième vérité, mais il faut faire preuve

le taux de 0% que le recensement laisse supposer et le taux de 4% qui est supposé pour le recensement. Dans le cas de sous-dénombrements globaux plus élevés pour les minorités, le recensement donne de meilleurs résultats que le redressement seulement quand il y a une gamme importante de variations entre les Etats et que les Etats avec une proportion relativement forte de minorités ont des taux de sous-dénombrement faibles. Par exemple, si le taux global est de 4%, le recensement donne une erreur de répartition de 0.8 comparativement à 1.2 pour le redressement, pourvu que le taux de sous-dénombrement pour tous les Etats avec une proportion relativement forte de minorités soit de 2% et de 6% pour tous les Etats avec une faible proportion de minorités. Si le taux global est de 5%, le recensement donne une erreur de répartition de 0.8 comparativement à 1.0, seulement si le taux de sous-dénombrement pour les Etats avec une proportion relativement forte de minorités est de 3% et de 7% pour les Etats avec une faibles proportion de minorités. Si le taux global est supérieur à 5%, il n'existe aucune combinaison de taux de sous-dénombrement dans les Etats où la gamme de variations est de 4% ou moins pour laquelle le recensement donne des résultats meilleurs que le redressement. L'étude du tableau 3 laisse supposer que les taux globaux devraient être de 3% ou moins pour que cette règle grossière du 5-1 donne des résultats moins précis que le recensement pour la répartition des sièges. Les 'intervalles de confiance' à 95% basés sur l'EP de 1990 pour le taux de minorité global vont de 4.3 à 5.7; cette gamme semble beaucoup trop étroite, mais même si nous doublons les marges d'erreur déjà mentionnées, l'intervalle irait de 3.6 à 6.4; si la valeur véritable se trouve entre ces limites, alors la règle du 5-1 donnera encore de meilleurs résultats que le recensement.

Redressements des chiffres du recensement basés sur le PEP et sur l'EP

Le programme de contrôle postcensitaire (PEP) de 1980 et l'enquête postcensitaire (EP) de 1990 visent à améliorer un redressement synthétique en estimant différents taux de sous-dénombrement dans différentes localités. Freedman et Navidi sont sceptiques à propos de la régression utilisée pour lisser les estimations, mettant en doute l'indépendance, l'homogénéité de la variance et la fiabilité des procédures de sélection utilisées pour inclure des variables dans le modèle. Les exemples qu'ils présentent de la façon dont on aurait facilement pu choisir différents ensembles de variables pour inclusion dans le modèle ne m'ont pas persuadé; après tout, s'il existe une forte corrélation entre les variables prédictives, des sous-ensembles fort différents peuvent produire à peu près la même prédiction. Ainsi, le fait que différentes variables aient été choisies ne signifie pas que les estimations lissées auraient été fort différentes. En effet, leur tableau 10 montre que deux variables, le pourcentage de minorités et la proportion de la population recensée selon la méthode classique figureraient dans presque toutes les équations. Je soupçonne que les *hypotheses* utilisées pour la régression ne peuvent être défendues facilement, mais que les *résultats* de la régression sont raisonnables, sauf peut-être qu'ils produisent des erreurs-types plus faibles que le manque probable d'indépendance ne le justifie.

La réduction de la variance d'échantillonnage à l'aide de procédures de lissage basées sur la régression ne fera probablement pas beaucoup de différence pour les estimations dans de grandes régions comme les Etats. Dans ces régions, le regroupement d'estimations différentes obtenues à partir du PEP ou de l'EP produit déjà tout le lissage nécessaire et l'on peut éviter les hypothèses discutables faites pour appliquer la régression. Par contre, le lissage à l'aide de la régression est probablement nécessaire si l'on veut projeter les résultats pour de petites localités. Je suis d'accord avec Freedman et Navidi pour dire que les procédures utilisées pour traiter le cas des données manquantes et l'évaluation du biais dans les enquêtes postcensitaires sont les éléments-clés pour évaluer les estimations redressées. Le traitement approprié des données manquantes et l'évaluation convenable du biais exigent une connaissance approfondie des procédures d'enquête. Les opinions personnelles des professionnels qui ont participé le plus étroitement au travail domineront les conclusions. On est justifié de faire preuve de scepticisme de bon aloi à propos de toutes les 'erreurs-types' ou de tous les 'intervalles de confiance' résultants.

Tableau 3
 Comparaison de l'erreur fractionnaire dans la répartition des sièges selon le recensement et selon un redressement 5-1 pour une gamme de taux de sous-dénombrement globaux pour les minorités, avec des taux de sous-dénombrement qui varient pour les États. (Le taux de sous-dénombrement pour la majorité est fixé à 1%; les minorités dans chaque État sont estimées à partir des données qui figurent dans l'ouvrage Statistical Abstract of the United States, 1989.)

Erreur fractionnaire dans la répartition des sièges selon un redressement 5-1	Erreur fractionnaire dans la répartition des sièges selon le recensement	Sous-dénombrement des minorités dans les États avec une proportion relativement forte de minorités	Sous-dénombrement des minorités dans les États avec une faible proportion de minorités	Sous-dénombrement global des minorités
0.7	0.2	2	2	2
1.1	0.4	1	3	2
0.6	0.7	3	1	2
0.5	0.5	3	3	3
0.9	0.4	2	4	3
0.4	0.9	4	2	3
0.2	0.7	4	4	4
0.7	0.6	3	5	4
0.4	1.1	5	3	4
1.2	0.8	2	6	4
0.9	1.6	6	2	4
0.0	0.9	5	5	5
0.5	0.8	4	6	5
0.5	1.3	6	4	5
1.0	0.8	3	7	5
1.0	1.8	7	3	5
0.2	1.2	6	6	6
0.4	1.0	5	7	6
0.7	1.6	7	5	6
0.9	1.0	4	8	6
1.2	2.0	8	4	6
0.5	1.4	7	7	7
0.4	1.2	6	8	7
0.9	1.8	8	6	7
0.8	1.2	5	9	7
1.4	2.2	9	5	7

de la façon appropriée. L'arrondissement fait de la répartition réelle un instrument de mesure plutôt mauvais pour comparer deux méthodes, parce que la différence dans les répartitions est habituellement de seulement 1 ou 2 sièges. J'utiliserai plutôt une erreur de répartition fractionnaire qui est la moitié de la somme des différences en valeur absolue entre les proportions estimées et véritables, multipliée par 435.

Les données qui figurent au tableau 3 nous permettent de constater que c'est lorsque le taux véritable de sous-dénombrement global pour les minorités est de 3% que les chiffres du recensement et les chiffres redressés donnent le même résultat; nous nous attendons à cela, parce qu'alors la véritable différence du taux de sous-dénombrement est de 2%, à mi-chemin entre

Tableau 1

Estimations historiques de l'importance et du pourcentage du sous-dénombrement net selon la race, telles que mesurées par l'analyse démographique. (Rapport du 21 juin 1991 du comité directeur sur le sous-dénombrement du Bureau of the Census.)

	1940	1950	1960	1970	1980	1990
Total	5.4	4.1	3.1	2.7	1.2	1.8
Noirs	8.4	7.5	6.6	6.5	4.5	5.7
Personnes non noires	5.0	3.8	2.7	2.2	0.8	1.3

Tableau 2

Sous-dénombrement estimé à partir de l'enquête postcensitaire (EP) et par l'analyse démographique (AD) pour le recensement de 1990, selon l'âge, la race et le sexe.

Noirs						Personnes non noires					
Hommes			Femmes			Hommes			Femmes		
EP	AD		EP	AD		EP	AD		EP	AD	
0-9	8.0	8.2	7.8	7.8	3.0	2.0	2.0	2.0	1.4	0.6	0.6
10-19	4.0	2.0	4.0	2.2	-1.3	3.3	2.7	3.4	1.8	2.8	-0.5
20-29	6.4	9.4	6.8	3.8	-0.3	5.0	2.1	3.8	1.4	0.9	0.1
30-44	5.9	12.4	3.9	2.5	-0.3	2.2	2.7	1.4	-0.5	0.4	0.4
45-64	3.2	11.7	1.3	0.5	-1.3	0.4	2.8	-0.5	-1.1	0.4	0.4
65 +	1.0	3.0	-0.3	-1.3	-0.9	1.4	1.4	-1.1	0.4	0.4	0.4

Je ne tiendrai pas compte des variations dans le sous-dénombrement des non-minorités entre les Etats, car ces variations devraient avoir un effet mineur sur les proportions globales; je supposerai que pour tous les Etats le sous-dénombrement des non-minorités est de 1%. Supposons que le véritable sous-dénombrement global des minorités soit de 5%. J'affecte les sous-dénombrements de 3% aux Etats avec une proportion relativement forte de minorité et les sous-dénombrements de 7% aux Etats avec une faible proportion de minorités, la division entre la proportion relativement forte et la faible proportion de minorités étant faite de façon à ce que le sous-dénombrement global des minorités soit de 5%. Cette hypothèse relative aux véritables sous-dénombrements fait le mieux paraître les résultats du recensement. Le calcul est effectué pour une gamme de choix du sous-dénombrement global et de variations entre les Etats. Pour comparer les chiffres du recensement et les estimations redressées, on calcule le nombre de sièges au Congrès qui sont mal répartis par les deux estimations de la population. Le nombre de sièges attribués à un Etat qui compte 7.2% de la population du pays est 435×7.2 arrondi

de minorités ont poursuivi le gouvernement. Le Secrétaire a alors accepté, le 17 juillet 1989, de poursuivre la planification de l'EP, de nommer un comité de huit experts qui le conseilleraient sur la faisabilité de redresser les chiffres du recensement et de publier un ensemble de lignes directrices qui permettraient de déterminer si les chiffres du recensement seraient ou non redressés. Le comité externe a tenu de nombreuses réunions avec des représentants du Bureau of the Census et leur a donné des conseils sur la planification, la réalisation et l'analyse de l'EP. Les analyses d'évaluation ont été effectuées par le Bureau et le 21 juin 1991, le comité directeur du recensement a recommandé, avec certaines dissidences, au Secrétaire que les chiffres du recensement soient redressés. La recommandation du comité a perdu beaucoup de poids quelques jours plus tard quand on a découvert qu'une analyse effectuée auparavant était erronée. Le comité externe s'est alors divisé en deux groupes de quatre personnes. Le premier de ces groupes, composé d'Ericksen, Estrada, Tukey et Wolter, avec l'aide de nombreux consultants, a écrit un rapport approfondi dans lequel on a trouvé de nombreuses lacunes relativement au dénombrement original et où l'on conseillait vivement de redresser les chiffres du recensement. Le deuxième groupe composé de Kruskal, McGeehee, Tarrance et Wachter, avec l'aide de quelques consultants, présente une autre analyse statistique de l'EP qui laisse supposer que la gamme de redressements possibles est tellement considérable qu'elle aurait des effets forts différents sur la révision du découpage électoral et sur d'autres exigences s'appliquant aux chiffres du recensement en matière de redistribution. Le Secrétaire a décidé que les bases statistiques d'un redressement étaient insuffisantes et il a recommandé de ne pas en effectuer. Le Département du Commerce a été poursuivi par les mêmes localités qui avaient intenté un procès en 1980. Le procès de 1980 est donc repris après le recensement de 1990.

Redressement synthétique

Un plan *synthétique* simple consiste à multiplier chaque personne, effectivement dénombrée, membre d'une minorité par 1.05 et chaque personne, effectivement dénombrée, de la majorité par 1.01. Je suis d'accord avec Freedman et Navidi pour le rejet des opinions exprimées dans

Schirm et Preston (1987).

Que faut-il penser de l'argument analytique suivant? Supposons que les sous-dénombrements nationaux soient estimés correctement, mais qu'ils diffèrent dans les États; quand le redressement synthétique des chiffres améliore-t-il la population proportionnelle d'un État? La réponse est, si le redressement synthétique consiste à augmenter les chiffres pour un État particulier, qu'il est plus proche de la proportion véritable que la proportion dénombrée si et seulement si la fraction des minorités dans l'État est inférieure à la fraction des minorités au niveau national; inversement, si le redressement consiste à diminuer les chiffres pour un État particulier, les chiffres redressés sont plus près de la proportion véritable que la proportion dénombrée si et seulement si la fraction des minorités dans l'État est supérieure à la fraction des minorités au niveau national. Il est plausible de s'attendre à ce que le sous-dénombrement des minorités et des non-minorités dans les États où l'on trouve une proportion relativement forte de minorités soit plus élevé que dans les États où l'on trouve une faible proportion de minorités, ce qui entraînerait un redressement synthétique causant un sous-dénombrement, mais qui serait néanmoins une amélioration par rapport à la proportion dénombrée. Les taux de sous-dénombrement à l'échelle nationale de 5% et de 1% sont appuyés par des données historiques du Bureau, obtenues tant à la suite d'analyses démographiques que d'enquêtes postcensitaires, tableaux 1 et 2. Quand on effectue un redressement 5-1, nous multiplions les chiffres du recensement pour les groupes non-minoritaires par 1.01 et les chiffres du recensement pour les groupes minoritaires par 1.05. Cette façon de calculer améliorera-t-elle la répartition des sièges des divers États au Congrès?

La population réelle des minorités et des non-minorités dans les différents États est inconnue. Nous comparons les deux estimations de la population d'un État basées sur les chiffres non redressés et redressés du recensement. Le recensement donnera les meilleurs résultats quand

COMMENTAIRES

J.A. HARTIGAN¹

La controverse relative au redressement

Toutes les dix années, le United States Census Bureau prépare une liste, ou un *dénombrement* des noms et des adresses des personnes résidant aux Etats-Unis. La liste est sujette à erreur en ce sens que des personnes peuvent ne pas y figurer ou y être incluses à tort. Erickson et coll. (1991) ont estimé que, pour le recensement de 1990, il y a eu 13 millions d'enregistrements erronés et 17 millions de personnes oubliées. Même si ces estimations sont deux fois plus grosses que les chiffres réels, il semble que les chiffres liés au dénombrement doivent subir un certain redressement.

Freedman et Navidi discutent de la preuve statistique présentée dans le cadre d'une poursuite intentée afin de forcer le Bureau of the Census à redresser les chiffres du recensement de 1980. C'est la *différence du taux de sous-dénombrement* entre les races qui est à l'origine de la poursuite. Le sous-dénombrement est peut-être de 5% pour les Noirs et les personnes d'origine hispanique et de 1% pour les autres personnes. Puisque le sous-dénombrement est plus élevé pour les minorités, les localités où l'on trouve de plus forts pourcentages de minorités font pression pour obtenir un redressement des chiffres du recensement qui corrigerait le sous-dénombrement. Ce dernier a été établi par analyse démographique (c'est-à-dire en comptant les naissances, les décès, l'émigration et l'immigration selon la race, le sexe et l'âge) dans les recensements depuis 1940 et par des enquêtes postcensitaires, afin d'obtenir un nombre plus exact dans un échantillon de la population, depuis 1970. L'importance du sous-dénombrement est un point important de discussion, puisqu'il est plus raisonnable d'estimer et de corriger une grosse différence du taux de sous-dénombrement qu'une petite. Freedman et Navidi admettent qu'une différence du taux de sous-dénombrement peut exister, mais ils soutiennent que, du moins pour le recensement de 1980, le sous-dénombrement n'est pas assez bien estimé dans différentes localités pour qu'on puisse effectuer un redressement. Freedman et Navidi sont surtout intéressés à critiquer les techniques proposées pour redresser les chiffres; quelles sont leurs propres estimations du sous-dénombrement? Par exemple, admettent-ils que le sous-dénombrement à l'échelle nationale est aussi élevé que 5% pour les Noirs et les personnes d'origine hispanique comparativement à 1% pour les autres habitants du pays? Je soutiendrai plus loin que si la différence du taux de sous-dénombrement est aussi élevée que ces chiffres le laissent supposer, un redressement synthétique (où l'on attribue à chaque personne membre d'une minorité la pondération 1.05 et à chaque personne membre de la majorité la pondération 1.01) améliorera probablement les estimations de la part de la population dans les Etats.

En 1980, le Bureau a appliqué un Post Enumeration Program (PEP) (programme de contrôle post-censitaire) dont il comptait utiliser les résultats pour redresser les chiffres du recensement. Le Bureau a décidé de ne pas procéder à un redressement des chiffres, parce que les estimations du PEP n'étaient pas suffisamment précises ou fiables pour produire des chiffres améliorés dans les petites localités. L'article reprend le témoignage de Freedman lors du procès qui a suivi, à la suite duquel le tribunal a rendu une décision favorable au Bureau of the Census. Il peut être intéressant de rapporter certains des faits qui se sont produits par la suite et qui montrent que les questions soulevées dans l'article dont nous traitons sont encore d'actualité. Au cours des années 80, le Bureau a préparé une enquête postcensitaire (EP) plus importante pour le recensement de 1990. Une répétition générale de l'EP a eu lieu en 1988. Environ 20 études d'évaluation portant sur divers types d'erreurs dans l'EP ont été préparées et effectuées après le recensement de 1990. En 1988, le Secrétaire au Commerce a annoncé que les chiffres du recensement de 1990 ne seraient pas redressés. Diverses localités comptant des proportions importantes

¹ J.A. Hartigan, Department of Statistics, Yale University, New Haven, CT USA 06520-2179.

F et N n'examinent pas minutieusement les procédures du recensement et les hypothèses sous-jacentes. Trouvent-ils plausible que l'"hypothèse" du recensement qui veut que lorsque les estimations finales sont diffusées, toutes les personnes, partout, ont été dénombrées exactement au bon endroit? Si ce n'est pas le cas, ont-ils des suggestions constructives à faire pour améliorer la précision des estimations démographiques? Malheureusement, leur commentaire critique sur les suggestions qui ont été proposées est gravement affaibli par la distorsion et par une présentation contenant des lacunes et il n'offre rien de constructif.

"Aurions-nous dû redresser les chiffres du recensement de 1980?" demandent F et N. Peut-être que oui, peut-être que non. Bien que cela puisse faire l'objet de discussions, il se peut que nous n'ayons pas eu suffisamment de renseignements sur les effets probables du redressement ou que nous n'ayons pas été en mesure, techniquement et opérationnellement, d'entreprendre un redressement au moment où la décision devait être prise. Le redressement aurait-il amélioré la précision en 1980? Nous ne pouvons répondre avec certitude parce que la population réelle est fondamentalement inconnaisable et qu'on ne peut écarter complètement la possibilité que des anomalies se produisent. Avec cette restriction, la réponse est "fort probablement".

Remerciements

Les auteurs remercient Gene Erickson qui leur a fourni des estimations non publiées provenant du Post Enumeration Program (PEP) de 1980. Les opinions exprimées sont celles des seuls auteurs.

SOURCES ADDITIONNELLES

ERICKSEN, E.P., et KADANE, J.B. (1983). Using the 1980 Census as a population standard. *Proceedings of the Social Statistics Section, American Statistical Association*, 474-479.

WOLTER, K.M. (1987). Comment on Schirm and Preston (1987). *Journal of the American Statistical Association*, 82, 978-980.

5. Discussion

améliore considérablement l'égalité de la représentation. D'après les chiffres réels de la population et la répartition effectuée en fonction des chiffres du recensement, il y a 471,000 personnes par représentant au Colorado, 508,000 dans l'Etat de New York, 555,000 en Géorgie et 558,000 en Alabama. Le redressement réduit les écarts, avec 565,000 personnes par représentant au Colorado, 524,000 dans l'Etat de New York, 505,000 en Géorgie et 489,000 en Alabama. Pour les quatre Etats combinés, il y a 519,000 personnes pour chacun des 57 représentants. (Pour tous les E.-U., il y a 519,000 personnes pour chacun des 435 représentants.) Le redressement réduit l'écart quadratique moyen (non pondéré), par rapport à cette moyenne, de plus de 21%. (La réduction est comprise entre 20 et 21% quand les écarts sont pondérés en fonction du nombre de personnes par représentant selon les chiffres réels de la population.) Le gain d'équité qui découle du redressement est aussi montré clairement par la moyenne pondérée du nombre de personnes par représentant calculée pour tous les 50 Etats. Quand on le pondère en fonction de la proportion de la population noire à l'échelle nationale (à l'exclusion du District de Columbia) vivant dans l'Etat, le nombre moyen réel de personnes par représentant est de 518,000. Si la répartition des sièges au Congrès est faite selon les estimations du recensement, la moyenne est de 524,000 personnes. Le redressement synthétique supprime la majorité de cette injustice basée sur la race. Il y a, en moyenne, 520,000 personnes par représentant quand les sièges de la Chambre sont répartis selon les estimations redressées. Bien qu'il existe certainement d'autres façons de mesurer l'inégalité de la représentation, il est difficile d'imaginer une autre méthode raisonnable qui ne montrerait pas que le redressement réduit l'injustice raciale attribuable à la différence du taux de sous-dénombrement dans le recensement. Le gain d'équité dans cet exemple est réalisé en dépit du fait qu'il n'y ait pas de gain d'exactitude dans la distribution proportionnelle entre les Etats.

Les erreurs dans le recensement sont systématiques. Après le redressement, il se peut que les erreurs restantes ne soient pas véritablement aléatoires et, tant qu'il y aura des erreurs, il y aura de l'injustice. Toutefois, la source de ces erreurs serait beaucoup moins offensante que la race.

Quand ils critiquent S et P et Erickson, Kadane et Tukey (1989), F et N soulignent le rôle des hypothèses qui sont à la base des estimations redressées. Leur opinion est cependant extrême, fondamentalement erronée et elle va à l'encontre du but recherché.

Bien qu'il soit raisonnable – et nécessaire – de se demander si les hypothèses ont de l'importance, il n'est pas nécessaire que ces dernières se vérifient exactement, comme F et N le laissent supposer. De plus, les partisans du redressement ne devraient avoir à défendre cette option que contre des hypothèses de remplacement raisonnables. F et N semblent croire que presque toute hypothèse de remplacement peut être utilisée. Comme lors de l'évaluation de l'importance de l'amélioration, ils exigent, sans justification scientifique, qu'un lourd fardeau de la preuve incombe aux partisans du redressement. Néanmoins, comme nos simulations le montrent, le redressement synthétique améliore la précision, même selon des hypothèses extrêmement défavorables – et probablement déraisonnables.

Cela soulève un autre point de désaccord important entre F et N et nous. Il n'est pas nécessaire que toutes les hypothèses mènent précisément aux mêmes estimations si toutes les estimations redressées basées sur des hypothèses raisonnables sont plus précises que les estimations du recensement. À moins que des hypothèses également plausibles aient des implications fort différentes, nous ne devrions pas rejeter le bien parce que nous n'avons pu trouver le mieux. Nous ne devrions pas nous contenter d'estimations du recensement qui sont moins précises. F et N ne présentent rien pour suggérer que les estimations du recensement sont plus précises que les estimations redressées. Les conclusions juridiques sur lesquelles ils s'appuient ne constituent pas une base pour une discussion scientifique. De plus, en dépit de la conclusion de Schirm (1991) que les décisions, basées sur des opinions, prises lors de la production des estimations du recensement peuvent avoir un effet sur la répartition des sièges au Congrès,

Tableau 2

Exemple numérique: Répartition des sièges au Congrès

Etat	Nombre de sièges à la Chambre des représentants		
	Réel	Recensement	Redressé
Alabama	8	7	8
Alaska	1	1	1
Arizona	5	5	5
Arkansas	4	4	4
Californie	46	45	45
Colorado	5	6	5
Connecticut	6	6	6
Delaware	1	1	1
District de Columbia	0	0	0
Floride	19	19	19
Géorgie	11	10	11
Hawaï	2	2	2
Idaho	2	2	2
Illinois	22	22	22
Indiana	10	10	10
Iowa	5	6	6
Kansas	5	5	5
Kentucky	7	7	7
Louisiane	8	8	8
Maine	2	2	2
Maryland	8	8	8
Massachusetts	10	11	11
Michigan	18	18	18
Minnesota	8	8	8
Mississippi	5	5	5
Missouri	9	9	9
Montana	2	2	2
Nebraska	3	3	3
Nevada	2	2	2
New Hampshire	2	2	2
New Jersey	14	14	14
Nouveau-Mexique	3	3	3
New York	33	34	33
Caroline du Nord	11	11	11
Dakota du Nord	1	1	1
Ohio	21	21	21
Oklahoma	6	6	6
Oregon	5	5	5
Pennsylvanie	23	23	23
Rhode Island	2	2	2
Caroline du Sud	6	6	6
Dakota du Sud	1	1	1
Tennessee	9	9	9
Texas	27	27	27
Utah	3	3	3
Vermont	1	1	1
Virginie	11	10	10
Washington	8	8	8
Virginie-Occidentale	4	4	4
Wisconsin	9	9	9
Wyoming	1	1	1
Total	435	435	435

Tableau 1
Exemple numérique: Chiffres de population (en milliers)

Etat	Réels		Recensement		Redressés	
	Noirs	Blancs	Noirs	Blancs	Noirs	Blancs
Alabama	1,060	2,849	996	2,898	1,051	2,866
Alaska	14	375	14	388	15	384
Arizona	83	2,606	75	2,643	79	2,614
Arkansas	414	1,916	374	1,912	395	1,891
Californie	1,758	22,105	1,819	21,849	1,919	21,607
Colorado	106	2,719	102	2,788	108	2,757
Connecticut	226	2,875	217	2,891	229	2,859
Delaware	101	507	96	498	101	492
District de Columbia	442	181	449	189	474	187
	1,415	8,290	1,343	8,403	1,417	8,310
Géorgie	1,642	3,910	1,465	3,998	1,546	3,954
Hawaii	18	925	17	948	18	938
Idaho	3	936	3	941	3	931
Illinois	2,014	9,677	1,675	9,752	1,768	9,644
Indiana	443	4,775	415	5,075	438	5,019
Iowa	43	2,809	42	2,872	44	2,840
Kansas	135	2,284	126	2,238	133	2,213
Kentucky	254	3,327	259	3,402	273	3,364
Louisiane	1,239	2,856	1,238	2,968	1,306	2,935
Maine	3	1,135	3	1,122	3	1,110
Maryland	1,026	3,159	958	3,259	1,011	3,223
Massachusetts	230	5,223	221	5,516	233	5,455
Michigan	1,272	7,990	1,199	8,063	1,265	7,974
Minnesota	56	4,037	53	4,023	56	3,978
Mississippi	891	1,629	887	1,634	936	1,616
Missouri	535	4,329	514	4,403	542	4,354
Montana	2	761	2	785	2	776
Nebraska	51	1,521	48	1,522	51	1,505
Nevada	52	760	51	749	54	741
New Hampshire	4	911	4	917	4	907
New Jersey	1,071	6,237	925	6,440	976	6,369
Nouveau-Mexique	27	1,264	24	1,279	25	1,265
New York	2,397	14,891	2,402	15,156	2,535	14,988
Caroline du Nord	1,387	4,443	1,319	4,563	1,392	4,512
	3	622	3	650	3	643
Ohio	1,112	9,769	1,077	9,721	1,136	9,613
Oklahoma	208	2,819	205	2,820	216	2,789
Oregon	39	2,602	37	2,596	39	2,567
Pennsylvanie	1,177	10,750	1,047	10,817	1,105	10,697
Rhode Island	29	923	28	919	30	909
Caroline du Sud	962	2,182	949	2,173	1,001	2,149
Dakota du Sud	2	678	2	689	2	681
Tennessee	754	3,764	726	3,865	766	3,822
Texas	1,752	12,421	1,710	12,519	1,804	12,380
Utah	10	1,435	9	1,452	9	1,436
Vermont	1	523	1	510	1	504
Virginie	1,116	4,356	1,009	4,338	1,065	4,290
Washington	109	3,867	106	4,026	112	3,981
Virginie-Occidentale	69	1,877	65	1,885	69	1,864
Wisconsin	198	4,588	183	4,523	193	4,473
Wyoming	3	448	3	467	3	462
Total	27,958	197,836	26,495	200,054	27,956	197,838

Nota: La catégorie "Blancs" comprend toutes les personnes non-noires.

À partir des estimations du PEP de 1980 publiées, nous pouvons seulement calculer les variances dans les taux de sous-dénombrement totaux pour les États, sans distinction selon la race. La plus importante variance entre les États parmi les 12 séries de la PEP publiée est 0.00034, légèrement moins que la variance moyenne simulée pour notre cas de variation moyenne (cas 22). Pour le cas 32 (faible variation dans le taux de sous-dénombrement des Noirs et variation moyenne dans le taux de sous-dénombrement des Blancs), la variance moyenne simulée est 0.00031, chiffre à peu près égal à la variance entre les États pour la série 2-9 du PEP, pour laquelle Erickson, Kadane et Tukey (1989) montrent une préférence et c'est aussi la variance médiane pour les huit séries du PEP qui restent après en avoir exclu les séries 10-8, 14-8, 14-9 et 14-20. Le redressement synthétique réduit la somme pondérée des carrés des erreurs d'environ 12% en moyenne pour le cas 32, comparativement à 8% pour le cas 22. Le cas 23 (variation moyenne du taux de sous-dénombrement des Noirs et variation faible du taux de sous-dénombrement des Blancs) laisse supposer une variance moyenne simulée qui n'est que légèrement supérieure à la variance pour la série 10-8 du PEP, la série "préférée" par F et N. Selon les hypothèses du cas 23, le redressement synthétique réduit la somme pondérée des carrés des erreurs de 19% en moyenne. Les estimations redressées sont plus précises dans plus de 92% des cas selon ce critère d'erreur. De telles améliorations sont-elles "seulement modestes"?

4. Précision et équité

Nous avons déjà insisté, dans S et P et dans Schirm (1991), sur le fait que les statisticiens et les démographes devraient se préoccuper, avant tout, de la précision des estimations de la population. Néanmoins, quand on consacre tous ses efforts à rechercher la précision statistique, il est facile d'oublier des considérations d'équité politique. Une distribution plus précise de la population est probablement plus équitable, en général. Cependant, cela ne signifie pas que deux distributions qui sont aussi précises, sont aussi équivalentes. Bien que le redressement puisse ne contribuer que très peu à l'amélioration de la précision globale au cours d'une année particulière, il peut réduire ou supprimer certaines erreurs et injustices systémiques, erreurs et injustices associées à la race. Un exemple, tiré de nos simulations, est donné au tableau 1. Les taux nationaux de sous-dénombrement implicites des Noirs et des Blancs sont de 5.2% et de -1.1%. Les estimations redressées de la population dans le tableau 1 ont été obtenues à l'aide de ces chiffres et de la méthode synthétique.

Comme cela deviendra clair, il est difficile de faire une distinction nette entre la précision et l'équité. Pour la présente discussion, nous supposons que la précision est définie de façon très stricte en fonction de la distribution géographique proportionnelle. Bien que les distributions selon les données du recensement et selon les chiffres redressés, présentées au tableau 1, soient aussi exactes, selon le critère de la somme pondérée des carrés des erreurs, les estimations redressées reflètent avec plus de précision la distribution raciale au niveau national et elles sont plus équitables. (La distribution géographique du recensement est un peu plus précise d'après la norme basée sur la somme de la valeur absolue des erreurs.) Les chiffres réels et redressés laissent supposer que les Noirs composent 12.4% de la population américaine. Selon les chiffres du recensement, les Noirs ne constituent que 11.7% de la population, une injustice considérable.

Les gains d'équité qui découlent du redressement sont rendus plus concrets par les répartitions implicites des sièges au Congrès présentées au tableau 2. Les estimations du recensement et les estimations redressées attribuent un siège de trop à l'Iowa et au Massachusetts et un siège de moins à la Californie et à la Virginie. Cependant, les estimations du recensement attribuent aussi un siège de trop au Colorado et à l'État de New York et un siège de moins à l'Alabama et à la Géorgie, alors que les estimations redressées attribuent le nombre approprié de sièges à ces États. L'Alabama et la Géorgie n'obtiennent pas le bon nombre de représentants en fonction des chiffres du recensement à cause de la proportion élevée de Noirs dans ces États et de la différence du taux de sous-dénombrement auquel les Noirs sont sujets. Le redressement

qu'elles renferment beaucoup de lacunes et qu'elles soient basées sur des hypothèses fragiles à propos de la migration interne (Wolter 1987), ces estimations laissent supposer un rapport direct entre le taux de sous-dénombrement des Noirs et le nombre de Blancs. Comme elles ne tiennent pas compte de l'une ou l'autre des structures de covariation, les simulations dans S et P tendent à sous-évaluer les gains de précision obtenus à l'aide du redressement synthétique.

Depuis que nous avons écrit S et P, nous avons obtenu des données non publiées, provenant du PEP de 1980, sur la population des Etats et sur les estimations du taux de sous-dénombrement selon la race. Parce que les estimations brutes du taux de sous-dénombrement des Noirs sont imprécises pour plusieurs Etats, il n'est pas clair si le dénombrement des Noirs est le plus difficile – au niveau de l'Etat – où ces derniers sont les plus nombreux. Pour les Blancs, bien qu'il existe une preuve d'une association directe, plutôt qu'inverse, entre leur nombre et le taux de sous-dénombrement, nous croyons que cela est attribuable à l'inclusion des personnes d'origine hispanique dans la population blanche et, dans une mesure beaucoup plus faible, au fait que l'on se fie beaucoup à la méthode de dénombrement conventionnelle dans quelques Etats, dont la population est surtout composée de Blancs, de l'Ouest des Etats-Unis. Nous trouvons, en effet, que si la véritable population de 1980 suivait l'évolution des estimations soit de la série 2-9, soit de la série 10-8, un redressement synthétique pour la différence entre le taux de sous-dénombrement des Noirs et des personnes d'origine hispanique et le taux de sous-dénombrement de toutes les autres personnes aurait presque certainement amélioré la précision des estimations.

Les preuves empiriques disponibles laissent généralement supposer que les variations géographiques sont, sinon aléatoires, systématiquement liées à une tendance qui améliorerait les gains de précision obtenus à l'aide du redressement synthétique. Il semble peu probable qu'il y ait une forte association inverse parmi les Etats entre le taux de sous-dénombrement des Noirs et le nombre de Blancs. (Même si l'une de ces tendances ou les deux existait, les estimations redressées pourraient encore être plus précises, comme nous l'avons montré dans S et P.) Ainsi, notre hypothèse du caractère aléatoire dans S et P était probablement prudente, défavorisant le redressement synthétique.

Nos hypothèses à propos de l'importance des variations du taux de sous-dénombrement entre les Etats étaient aussi probablement prudentes, comme nous en avons traité dans S et P en nous reportant aux *Developmental Estimates* de 1970. A cause d'erreurs d'échantillonnage importantes pour de nombreux Etats, les estimations non publiées du taux de sous-dénombrement dans les Etats selon la race, provenant du PEP de 1980, ne révèlent pas de façon fiable comment le taux de sous-dénombrement des Noirs varie entre les Etats. La variance dans le taux de sous-dénombrement des Noirs calculée pour tous les 51 Etats est 0.0128 pour la série 2-9, deux fois la valeur la plus élevée supposée dans nos simulations. La variance tombe à 0.0036, valeur qui n'est même pas à mi-chemin entre les variances moyenne et élevée simulées, quand on exclut le New Hampshire (taux de sous-dénombrement des Noirs égal à – 60%) et le Vermont (taux de sous-dénombrement des Noirs égal à – 24%). (Pour la série 10-8, la variance entre les Etats est presque égale à la valeur moyenne simulée dans S et P, si trois Etats avec un taux de sous-dénombrement extrême (et très peu fiable) – inférieur à – 20% ou supérieur à 20% – sont exclus). Les estimations brutes, provenant de la PEP de 1980, du taux de sous-dénombrement des Blancs dans les Etats sont beaucoup plus précises. Les variances entre les Etats pour les séries 2-9 et 10-8 sont très légèrement inférieures à la valeur moyenne simulée. Les gains au niveau de la précision d'après le scénario I, cas 12 dans S et P (variation élevée dans le taux de sous-dénombrement des Noirs et variation moyenne dans le taux de sous-dénombrement des Blancs) diffèrent peu en fréquence ou en importance par rapport aux gains obtenus selon le cas 22 (variation moyenne pour les taux de sous-dénombrement des Noirs et des Blancs), où des améliorations sont fort probables.

Nous soupçonnons que certaines de ces procédures ne donneraient pas de trop bons résultats, ayant presque certainement fait empirer les différences dans le taux de sous-dénombrement plutôt que de les avoir améliorées. Finalement, le redressement serait-il recommandé s'il améliorerait très peu la précision mais s'il réduisait l'injustice systématique? Nous reprendrions cette dernière question dans la Section 4.

F et N ne répondent qu'implicitement, si même ils le font, à ces questions qui, de façon générale, n'ont pas de réponses statistiques. Ils suggèrent, cependant, que le redressement pourrait être intéressant (ses estimations "seront bonnes") si les hypothèses présentées dans notre article se vérifient, question à laquelle nous passons maintenant.

F et N caractérisent incorrectement à la fois les méthodes synthétiques et notre modèle de simulation. L'hypothèse qui est à la base de la méthode synthétique n'est pas qu'il n'existe pas de variation géographique systématique dans le taux de sous-dénombrement pour une race donnée, mais qu'il n'y a aucune variation. Notre modèle de simulation montre comment le redressement synthétique fonctionne quand cette hypothèse à propos de la méthode synthétique n'est pas respectée. Nous avons considéré des cas extrêmes, bien que non systématiques, de variation entre les Etats du taux de sous-dénombrement selon la race, ainsi que des cas avec une variation aléatoire plus modérée. Nous n'avons pas construit de population "réelle", "suivant l'hypothèse de la méthode synthétique", et notre "définition de la 'réalité'" ne "favorise pas le redressement synthétique". Comme nous l'avons montré de façon analytique dans S et P, le redressement synthétique aurait été favorisé si l'on avait supposé une association positive entre le taux de sous-dénombrement des Noirs et le nombre de Noirs ou une association négative entre le taux de sous-dénombrement des Blancs et le nombre de Blancs. (Un énoncé précis du résultat est présenté dans l'annexe à S et P.)

Seulement à des fins d'illustration, nous avons supposé dans une nouvelle série de simulations que le taux de sous-dénombrement des Blancs est produit selon les hypothèses du scénario 1, cas 22 présenté dans S et P, mais que le taux de sous-dénombrement prévu pour les Noirs augmente avec la proportion de Noirs dans l'Etat de sorte que le taux de sous-dénombrement des Noirs est de 2.0% quand la proportion de Noirs est de 11.7% (la proportion de Noirs à l'échelle nationale en 1980 selon le recensement) et de 5.2% quand la proportion de Noirs est de 20.0%. Dans ces conditions, que nous ne prétendons pas être réalistes, bien qu'elles pré-servent la différence moyenne simulée du taux de sous-dénombrement national, le redressement synthétique améliore la précision de la distribution proportionnelle selon le critère de la somme pondérée des carrés des erreurs dans toutes les 1,000 itérations. La réduction moyenne de la somme pondérée des carrés des erreurs dépasse 17%, en dépit de la variation extrême dans le taux de sous-dénombrement total pour les Etats.

Les hypothèses présentées dans S et P à propos des variations du taux de sous-dénombrement entre les Etats se vérifient-elles? Probablement pas. Bien qu'un de nos buts était de simuler une gamme étendue de circonstances, il est fort probable que nos hypothèses avaient tendance à mettre le redressement dans une position désavantageuse.

Avons-nous apporté "aucune réponse" au sujet de la variation géographique, comme F et N le prétendent? Non, bien qu'il faille reconnaître qu'on ne disposait pas d'une abondance de renseignements. Il existe deux questions empiriques pertinentes pour juger nos hypothèses et leurs implications: la variation est-elle systématique ou aléatoire? et est-elle importante? Dans S et P, nous avons abordé ces deux questions.

Comme nous l'avons fait remarquer dans S et P, selon le PEP de 1980, c'est quand ils consistent une proportion importante de la population qu'il est le plus difficile de dénombrer les Noirs. Par contre, il n'y a essentiellement aucun rapport entre le taux de sous-dénombrement des Blancs et le nombre relatif de Blancs. (Ericksen et Kadane 1983). Ces conclusions sont basées sur de grandes catégories utilisées pour mesurer la composition raciale et sur des données pour les Standard Metropolitan Areas (SMSA) ainsi que pour les Etats amputés de leurs SMSA, et non sur des données au niveau de l'Etat. Les seules estimations publiées du sous-dénombrement selon l'Etat et la race sont les "Developmental Estimates" pour 1970. Bien

desquelles dépend le succès ou l'échec du redressement. Selon les preuves empiriques disponibles citées plus loin, les conditions qui conviennent pour l'exemple numérique de F et N ne s'appliquaient pas en 1980 et nos résultats laissent supposer que le redressement synthétique aurait amélioré la précision de la distribution géographique.

3. Résultats des simulations

Comme nous l'avons déjà fait remarquer, l'objet de nos simulations était de répondre à plusieurs questions relatives au redressement synthétique et à ses effets sur la précision des estimations démographiques. Voici quelles étaient les questions centrales abordées dans notre article:

- Dans quelle proportion des cas le redressement synthétique améliorerait-il la précision des estimations démographiques?
- Dans quelle mesure le redressement synthétique améliorerait-il généralement la précision des estimations démographiques?
- Les effets du redressement synthétique sur la précision dépendent-ils de la mesure dans laquelle la couverture du recensement varie d'un Etat à l'autre?
- Les effets du redressement synthétique sur la précision dépendent-ils de la façon dont nous pouvons bien mesurer le taux de sous-dénombrement national?

F et N se concentrent sur la deuxième question. Dans l'ensemble, nous admettons que l'importance moyenne de l'amélioration de la précision résultant du redressement synthétique est modeste si nos hypothèses modérées à propos de l'état de la nation sont justifiées. D'après le cas 22 du scénario I, qui exagère probablement les variations de la couverture du recensement entre les Etats mais qui est présente dans S et F comme notre cas de variation "modéré", la réduction moyenne de la somme pondérée des carrés des erreurs n'est que de 8% alors que la réduction moyenne de la somme non pondérée de la valeur absolue des erreurs n'est que de 4%. Il est, cependant, important de comprendre que des améliorations plus considérables pourraient être réalisées, comme notre troisième résultat analytique présente dans S et F le laisse supposer. Les gains au niveau de la précision seraient un peu plus considérables, par exemple, si le taux de sous-dénombrement national des personnes d'origine hispanique était identique à celui des Noirs et si ces personnes étaient incluses avec les Noirs plutôt qu'avec les Blancs. Dans ce cas, la réduction moyenne de la somme pondérée des carrés des erreurs serait supérieure à 12%. Les gains au niveau de la précision seraient aussi plus considérables si le taux de sous-dénombrement des Noirs était plus élevé dans les Etats comptant proportionnellement plus de Noirs. Nous reviendrons à cette question sous peu. Bien entendu, les améliorations obtenues à la suite d'un redressement synthétique pourraient être plus faibles si des erreurs considérables s'étaient produites dans la mesure du taux de sous-dénombrement, bien que, comme nous l'avons montré dans S et F, les effets des erreurs de mesure sont généralement faibles.

Quand on évalue l'amélioration moyenne de la précision, on oublie facilement la probabilité d'obtenir une amélioration, grande ou petite. F et N ont fait cette erreur. Selon les hypothèses du cas 22, scénario I, la probabilité d'améliorer la précision, d'après le critère de la somme pondérée des carrés des erreurs, est de 84%. Nous sommes impressionnés par cette constatation. Il est fort probable qu'une certaine amélioration, peut-être seulement modeste, sera réalisée.

Ce résultat ainsi que le résultat relatif à l'importance moyenne de l'amélioration soulèvent des questions très importantes. Qu'implique le fait que l'amélioration moyenne soit "seulement modeste"? L'amélioration moyenne doit-elle être écrasante pour justifier un redressement? Ou pour nous exprimer autrement, les estimations redressées devraient-elles être sujettes à une norme plus stricte que les estimations du recensement? Le Secrétaire au Commerce a imposé une norme plus stricte lorsqu'il a pris sa décision à propos du redressement des chiffres du recensement de 1990. Comment les procédures d'amélioration de la couverture et d'imputation du Census Bureau se comporteraient-elles si elles étaient soumises à une norme aussi stricte?

viole les conditions que nous avons élaborées et énoncées précédemment dans l'annexe à notre article. Pour répéter ce résultat, nous avons montré que le redressement peut rendre la proportion estimée de la population nationale totale résidant dans l'Etat i plus précise si

$$\left| \sum_{f=1}^J \frac{N_C^f}{N_T^f} \left(\frac{N_C^f}{N_T^f} \right) - \sum_{f=1}^J \frac{N_C^f}{N_T^f} \left(\frac{N_C^f}{N_T^f} \right) \right| > \left| \sum_{f=1}^J \frac{N_C^f}{N_T^f} \left(\frac{N_C^f}{N_T^f} \right) - \sum_{f=1}^J \frac{N_C^f}{N_T^f} \left(\frac{N_C^f}{N_T^f} \right) \right|$$

où J représente le nombre de groupes raciaux (plus généralement, de groupes démographiques), un point indique qu'il faut faire la somme pour toutes les valeurs d'un indice et les exposants T et C désignent les chiffres de population réels et obtenus dans le cadre du recensement, respectivement. Pour l'Etat A dans le "contre-exemple" de F et N, cette expression donne

$$\left| \frac{90}{1,100} \left(\frac{1,100}{1,100} - \frac{90}{89} \right) + \frac{1}{1,100} \left(\frac{1,100}{1,100} - \frac{1}{2} \right) \right| = 0$$

$$\neq \frac{21}{13,750} = \left| \frac{90}{1,100} \left(\frac{1,100}{979} - \frac{90}{89} \right) + \frac{1}{1,100} \left(\frac{1,100}{121} - \frac{1}{2} \right) \right|$$

Par conséquent, la condition pour que le redressement améliore la précision est violée pour l'Etat A. De même, pour l'Etat B, nous obtenons

$$\left| \frac{910}{1,100} \left(\frac{1,100}{890} - \frac{910}{910} \right) + \frac{1,100}{99} \left(\frac{1,100}{1,100} - \frac{99}{119} \right) \right| = 0$$

$$\neq \frac{21}{13,750} = \left| \frac{910}{1,100} \left(\frac{1,100}{979} - \frac{910}{890} \right) + \frac{1,100}{99} \left(\frac{1,100}{121} - \frac{99}{119} \right) \right|$$

A nouveau, la condition pour que le redressement améliore la précision est violée.

Le "contre-exemple" de F et N ne dit rien à propos de notre deuxième résultat analytique. Toutefois, il est utile pour illustrer numériquement notre troisième résultat qui est manifestement le plus important. Selon ce résultat, quand le taux de sous-dénombrement des Noirs est le plus important, dans les Etats où ces derniers sont *les moins* nombreux et le taux de sous-dénombrement des Blancs est le plus élevé dans les Etats où ces derniers sont *les plus* nombreux, *il se peut* que le redressement synthétique n'améliore pas la précision de la distribution proportionnelle. Dans l'exemple de F et N, le taux de sous-dénombrement des Noirs dans l'Etat A est plus élevé que dans l'Etat B (50% par opposition à 17%) mais l'Etat A compte proportionnellement moins de Noirs que l'Etat B (2% par opposition à 12%). Le taux de sous-dénombrement des Blancs est plus élevé dans l'Etat A (plus petit surdénombrement) que dans l'Etat B (-1% par opposition à -2%) et l'Etat A renferme proportionnellement plus de Blancs (98% par opposition à 88%). Par conséquent, bien que cela ne soit pas assuré, la conclusion de F et N que les estimations redressées dans leur exemple sont moins précises dans l'ensemble que les estimations du recensement n'est pas surprenante compte tenu de notre troisième résultat analytique.

La critique que font F et N de notre analyse algébrique des effets du redressement est basée sur une lecture très sélective de notre article qui présente de manière incorrecte nos conclusions. La critique que font F et N de notre deuxième résultat analytique est erronée tout comme le fait qu'ils caractérisent ce résultat comme central pour notre article. Notre troisième résultat analytique est, de beaucoup, plus important. Ce dernier nous aide à exposer les conditions

COMMENTAIRES

ALLEN L. SCHIRM et SAMUEL H. PRESTON¹

1. Introduction

Nous remercions le rédacteur en chef qui nous a invités à faire des commentaires sur cet article sujet à controverse de Freedman et Navidi (que nous désignerons dorénavant par "F et N") et à continuer cet important débat en matière de politique. Nos commentaires visent surtout à répondre aux critiques faites par F et N à propos de notre recherche antérieure (Schirm et Preston, 1987; désigné dorénavant par "S et P"). Bien que nous ne soyons pas d'accord avec une bonne partie de la critique que font F et N de Ericksen, Kadane et Tukey (1989), nous laissons aux auteurs de ce dernier article la tâche de défendre leur travail.

Nous ne sommes pas d'accord avec un bon nombre des critiques précises que font F et N de S et P. Avant d'examiner nos réponses détaillées, nous désirons considérer les choses d'une façon plus générale et voir notre article ainsi que les critiques de F et N dans leur ensemble. F et N caractérisent mal notre article en disant que nous présentons "deux arguments majeurs: l'un analytique et l'autre fondé sur la simulation". En fait, nous avons présenté trois résultats analytiques. F et N n'en critiquent qu'un et encore il s'agit d'un argument mineur. Notre résultat analytique le plus important laisse supposer que le redressement synthétique aurait vraisemblablement amélioré l'exactitude de la distribution de la population en 1980. Pour ce qui est de nos simulations, elles n'avaient pas pour but d'appuyer un argument quelconque. Nous avons plutôt simulé une gamme extrêmement étendue de circonstances afin d'être en mesure de nous attaquer à plusieurs questions portant sur le redressement synthétique et sur ses effets. Nous avons cependant trouvé que le redressement aurait amélioré la précision dans toutes les conditions simulées, y compris dans des circonstances très défavorables.

2. Résultats analytiques

Dans S et P, nous avons présenté trois résultats analytiques. Tous les trois sont mathématiquement exacts. Toutefois, le deuxième résultat – le seul qui fasse l'objet des critiques de F et N – pourrait, comme nous l'avons mentionné dans notre article, "induire en erreur parce qu'il ne tient pas compte des influences, sur le succès du redressement global, des rapports systématiques entre les variations d'un Etat à l'autre dans la couverture du recensement pour un groupé et des différences, entre les groupes, dans la façon dont ces derniers sont distribués entre les Etats. (TRADUCTION)" Notre troisième résultat, qui est manifestement au centre de notre analyse algébrique et qui ne dépend pas du deuxième résultat, s'attaque à cette question et tient compte des tendances dans les variations du taux de sous-dénombrement entre les Etats. Bien qu'il puisse induire en erreur, nous avons présenté le deuxième résultat pour illustrer avec plus de vigueur une implication fondamentale de notre troisième résultat, soit le fait que les variations systématiques dans le taux de sous-dénombrement au niveau des Etats peuvent avoir de l'importance.

Notre deuxième résultat analytique laisse supposer que l'effet du redressement pour un Etat donné dépend de la mesure dans laquelle le taux de sous-dénombrement pour l'Etat "se rapproche" du taux de sous-dénombrement national. Contrairement à ce que F et N déclarent, notre deuxième résultat analytique est mathématiquement exact. C'est seulement parce qu'ils choisissent de définir "se rapproche" sans tenir compte de notre définition précise que F et N sont en mesure de mettre nos conclusions en doute. Ainsi, le "contre-exemple" de F et N à notre deuxième résultat ne se rapporte pas du tout à ce résultat, puisque leur exemple

¹ Allen L. Schirm, Mathematica Policy Research, Inc., 600 Maryland Avenue, S.W., Suite 550, Washington, D.C. USA 20024-2512, Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA USA 19104-6298.

technologie de pointe. L'enquête post-censitaire de 1990, bien conçue, bien appliquée et avec une bonne assurance de la qualité et un excellent appariement informatique ainsi que des éléments géographiques précis nous fait voir la situation de 1980 de façon fort différente. Selon moi, le redressement peut maintenant être réalisé avec succès au niveau des Etats. La recherche et la discussion afin de savoir si ce redressement réussit peut aussi être effectuée à des niveaux géographiques inférieurs méritent qu'on leur consacre nos ressources collectives (p. ex., Tukey 1983; Cressie 1988; Wolter et Causey 1991). On peut utiliser les pertes (ou les risques) *prévus* pour mesurer l'efficacité des procédures de redressement. Cressie (1988) donne des conditions suffisantes qui, lorsqu'elles sont respectées, assurent que les résultats du redressement synthétique sont meilleurs que les chiffres du recensement; ces conditions étaient satisfaites lors du recensement de 1980 et dans la série PEP 3-8.

SOURCES ADDITIONNELLES

CHILDBERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., et WOLTER, K. (1987). Document présenté au session *1990 Census Undercount and Adjustment*, American Statistical Association Annual Meetings, San Francisco, California.

CRESSIE, N. (1988). Dans quelles circonstances les opérations de redressement améliorent-elles les chiffres du recensement? *Techniques d'enquête*, 14, 205-222.

CRESSIE N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

CRESSIE, N. (1990). Weighted smoothing of estimated undercount. *Proceedings of Bureau of the Census 1990 Annual Research Conference*. Bureau of the Census, Washington, D.C., 301-325.

CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

PASSELL, J.S., et ROBINSON, J.G. (1984). Revised estimates of the coverage of the population in the 1980 Census based on demographic analysis: A report on work in progress, dans *Proceedings of the Social Statistics Section, American Statistical Association*, 160-165.

PASSELL, J.S., SIEGEL, J.S., et ROBINSON, J.G. (1982). *Coverage of the National Population in the 1980 Census, by Age, Sex and Race: Preliminary Estimates by Demographic Analysis*. Current Population Reports, Special Studies P-23, No. 115. Bureau of the Census, Washington, D.C.

PRASAD, N.G.N., et RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

SIEGEL, J.S. (1974). Estimates of the coverage of the population by sex, race and age in the 1970 Census. *Demography*, 11, 1-23.

TUKEY, J.W. (1983). Affidavit présenté au District Court, Southern district of New York. *Cuomo et al. versus Baldridge*. 80 Civ. 4550 (JES).

recensement de 1990 pour éliminer les différences dans le taux de sous-dénumbrement" (TRADUCTION): Childers et coll. 1987).

Il est temps que Freedman et Navidi abandonnent leur rôle d'avocats du diable; il est temps qu'ils utilisent leurs connaissances et leurs talents de façon constructive et il est temps qu'ils disent ce qu'ils entendent par "assez de précision", "niveau local", et par diverses autres affirmations d'ordre qualitatif. La polémique engendrée par les débats judiciaires a gagné les divers articles, commentaires et répliques publiés sur le sous-dénumbrement du recensement au cours des dix dernières années. Pour résoudre un problème aussi difficile que le redressement des chiffres en fonction du sous-dénumbrement, il faut reconnaître le but commun. Une fois que cela est fait, la discussion devrait se concentrer sur les différences dans les moyens qui pourraient permettre d'atteindre ce but. Si Freedman et Navidi pensent que cela est impossible (ce qu'ils semblent avoir laissé supposer au fil des ans), alors il faudrait le dire explicitement.

Dans le reste du présent commentaire, je vais m'attaquer à un certain nombre de questions techniques importantes qui ont été soulevées par Freedman et Navidi (1986) et qui, c'est surprenant, le sont à nouveau dans l'article sur lequel porte le présent commentaire. En 1990, j'ai présenté lors de la Annual Research Conference du Census Bureau (Cressie 1990) une communication que David Freedman a été invité à commenter. À la dernière minute, il n'a pu assister à la conférence mais j'ai continué à lui envoyer la version de travail et la version finale de la communication et je l'ai invité à faire ses commentaires. La communication est plutôt technique mais elle s'attaque, avec succès je le crois, à plusieurs des critiques importantes faites par Freedman et Navidi (1986) sur la méthode de modélisation statistique pour redresser les chiffres en fonction du sous-dénumbrement.

Premièrement, dans la communication j'ai exprimé une préférence pour la "méthode de stratification" plutôt que pour la "méthode de régression". La stratification est un cas spécial de la régression où la valeur des variables explicatives est limitée à 1 et à 0, pour indiquer la présence ou l'absence dans une strate (démographique) particulière. Presque tout le monde admet qu'il y a une différence dans le taux de sous-dénumbrement à travers les strates sexe \times âge \times race/origine ethnique. Parce que le Census Bureau ne pensait qu'à utiliser une méthode de régression, la majorité de l'article s'attaquait au problème plus général.

Deuxièmement, on pourrait permettre que l'erreur de régression (voir Freedman et Navidi ex. (2)) soit dépendante; dans de tels modèles, le terme d'erreur peut inclure le biais et les erreurs de spécification. Le concept important qui doit être conservé à l'esprit est que le véritable sous-dénumbrement dans les régions est inconnu et que cette ignorance est quantifiée dans un modèle probabiliste. Le but n'est pas d'estimer les coefficients β mais de prédire le sous-dénumbrement. Avec un terme d'erreur qui n'a pas à être indépendant et distribué de façon identique, cette prévision n'est pas sensible aux erreurs de spécification (voir aussi Cressie 1991, chapitre 3). Troisièmement, on s'attaque à l'inconsistance du modèle relativement aux changements dans le niveau géographique en modélisant les facteurs de redressement, et non les sous-dénumbrements, et en supposant que la variance de l'erreur de régression pour une région particulière est inversement proportionnelle à la population de cette région. Cette hypothèse est justifiée, d'un point de vue tant bayésien que "fréquentiste", dans Cressie (1989).

Quatrièmement, on peut tenir compte de l'effet de l'estimation des paramètres de variance-covariance en modifiant les résultats de Prasad et Rao (1990) pour les adapter à un contexte produisant des données à l'aide du modèle estimé, puis en ré-estimant tous les paramètres et finalement en faisant une nouvelle prévision du sous-dénumbrement.

Finalement, on admet que toutes les méthodes à base de modèles décrites plus haut donneront vraisemblablement de mauvais résultats si le modèle est mal ajusté. Les méthodes de diagnostic ont une importance critique pour le succès des redressements, basés sur des modèles statistiques, pour tenir compte du sous-dénumbrement.

Il y a lieu de faire une évaluation critique de nos succès et de nos échecs passés. Il est temps d'aller de l'avant et de résoudre ce problème, qui a une importance monumentale, avec une

COMMENTAIRES

N. CRESSIE¹

L'évaluation critique de nos succès et de nos échecs passés nous prépare mieux à assurer d'autres succès. Les problèmes relatifs aux données manquantes et à l'appariement dans le Post Enumeration Program (programme de contrôle postcensitaire) de 1980 ont été des obstacles importants à un redressement réussi des chiffres du recensement décennal des États-Unis de 1980. Un procès (*Cuomo et al. v. Baldridge*) a été intenté par l'État de New York et d'autres parties afin d'obliger le Census Bureau à redresser les chiffres du recensement de 1980 pour tenir compte du sous-dénombrement. Les témoignages de Barbara Ballar, alors directrice adjointe, Statistical Standards and Methodology, du Census Bureau et de Kirk Wolter, alors chef de la Statistical Research Division du Census Bureau, ont démontré clairement que les données et les méthodes du recensement de 1980 étaient inadéquates pour redresser avec précision les chiffres pour l'ensemble du pays.

En 1987, le juge Spizzo s'est prononcé contre l'État de New York. Toutefois, cette décision n'a pas fait disparaître la différence du taux de sous-dénombrement; même le juge, dans sa décision, a reconnu sa présence. Presque tous les intéressés admettent que, à des taux différents selon la race, les chiffres nationaux du recensement des E.-U. ont été constamment trop faibles. Les méthodes démographiques nous permettent de fournir les estimations suivantes:

1950: Le sous-dénombrement, estimé par des méthodes démographiques, des Noirs (et des autres groupes de personnes non blanches) était de 9.7%, et pour les Blancs de 2.5%. (Siegel 1974, tableau 3).

1960: Le sous-dénombrement, estimé par des méthodes démographiques, pour les Noirs (seulement) était de 8.0%, et pour les Blancs (et les autres races) de 2.1%. (Siegel 1974, tableau 2, estimations de l'ensemble D).

1970: Le sous-dénombrement, estimé par des méthodes démographiques, pour les Noirs (seulement) était de 7.6%, et pour les Blancs (et les autres races) de 1.5%. (Passel, Siegel et Robinson 1982, tableau 1).

1980: Le sous-dénombrement, estimé par des méthodes démographiques, pour les Noirs (seulement) était de 5.3%, et pour les Blancs (et les autres races) de - 0.2%. (Passel et Robinson 1984, tableau 2).

De plus, presque tous les intéressés conviennent que la composition raciale diffère dans les régions administratives (qu'elles soient grandes ou petites) dans tous les E.-U. La conséquence de ces deux faits virtuellement indéniables est que le taux de sous-dénombrement variera selon les régions administratives, ce qui aboutira à la production d'un profil géographique/racial non représentatif de la nation et à une répartition injuste des ressources politiques et financières. Freedman et Navidi déclarent donc dans leur introduction "... Si le sous-dénombrement peut être estimé avec assez de précision, surtout au niveau local, on peut envisager - et même recommander - des opérations de redressement en vue d'améliorer les chiffres du recensement." Presque tout le monde admet qu'il y a un problème. L'adage, "le mieux est l'ennemi du bien", ne s'applique pas ici. Un professionnel de la statistique se trouve dans une position défensive inconfortable quand il soutient que les biais et les erreurs dont on n'a pas tenu compte ne permettront pas d'effectuer un redressement des chiffres dans le cas d'un sous-dénombrement dont on connaît l'existence et dont on sait qu'il est nuisible. Au cours du début des années 80, Ballar et Wolter ont créé un Undercount Research Staff au sein de la Statistical Research Division du Census Bureau. Les membres de ce service ont produit de la recherche de haute qualité qui a démontré "qu'il est techniquement possible de corriger les chiffres du

¹ N. Cressie, Département de statistique, Iowa State University, Ames, Iowa E.-U. 50011.

- a) Les chiffres du recensement forment une base de données multidimensionnelle intégrée. La nécessité de redresser ces chiffres reste à démontrer, même si (et encore là) la variable la plus élémentaire – le chiffre de population – peut être améliorée grâce à cette opération.
- b) Si, de fait, il est possible de produire un ensemble d'estimations démographiques que l'on juge supérieures aux chiffres du recensement (selon des critères appropriés), produisons-les et utilisons-les sans nécessairement redresser toute la base de données du recensement. Les critères devraient avoir rapport à l'ensemble de régions (ou autres unités géographiques) pour lesquelles on calcule des estimations.
- c) Si la loi exige l'utilisation de chiffres de population tirés du recensement alors qu'en réalité, il nous faut les meilleures estimations démographiques qui puissent exister, il semblerait préférable d'essayer de modifier la loi au lieu de redresser (pondérer en fait) toute la base de données du recensement de manière à ce qu'elle concorde avec les estimations démographiques dans le seul but de dire que ces estimations **sont** les chiffres du recensement.
- d) Les considérations d'équité, tant pour la répartition des fonds publics que pour celle de la représentation politique, sont valables pour toute la période intercensitaire et non seulement pour l'année du recensement. On devrait les analyser à l'aide de modèles qui tiennent parfaitement compte de ce fait.

3. Conclusion

(1) Les formules exploitent une diversité de renseignements statistiques (dont la plupart viennent d'autres sources que le recensement) et les chiffres de population ne constituent qu'une partie de ces renseignements. Par ailleurs, on sait très bien que pour plusieurs autres composantes, l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage sont appréciables. On ignore encore s'il est possible d'améliorer réellement toute fonction de perte acceptable destinée à évaluer l'effet combiné de toutes les erreurs, même en supposant que les erreurs du recensement soient entièrement éliminées.

(2) Il y a un point plus important encore. Si les chiffres de population redressés produisent une fonction de perte moindre ou si, de façon plus générale, on trouve qu'ils reflètent plus la réalité pour une grande majorité de régions, on peut alors s'en servir pour produire de meilleures **estimations démographiques** (et non seulement dans les années de recensement) sans pour autant redresser toute la base de données du recensement. Au Canada (ainsi qu'aux États-Unis), on publie depuis longtemps des estimations du sous-dénombrement dans le recensement. On envisage sérieusement maintenant, au Canada, de passer à l'étape suivante: d'une part, publier les résultats du recensement tels quels et d'autre part, produire une série d'estimations démographiques officielles qui tient compte des chiffres connus du sous-dénombrement. Après tout, les estimations démographiques officielles produites dans les années intercensitaires sont établies au moyen de toutes sortes de méthodes, dont certaines comportent des erreurs au moins aussi grandes que celles contenues vraisemblablement dans les estimations du sous-dénombrement (même les estimations de modèle). D'un point de vue scientifique, il peut être tout à fait convenable de publier les meilleures estimations démographiques dont on dispose aussi bien dans les années de recensement que dans les années intercensitaires, peu importe que ces estimations soient conformes ou non aux chiffres du recensement dans une année de recensement. Il faudra peut-être modifier la loi, ou les règlements pertinents, pour permettre l'utilisation – particulièrement l'année d'un recensement – d'estimations démographiques différentes de celles tirées directement du recensement. Toutefois, a) cela n'a pas grand-chose à voir avec les arguments scientifiques qui ont été présentés dans le débat sur le redressement et b) cette façon de procéder est plus honnête que d'identifier les chiffres "redressés" comme les résultats officiels du recensement sous le simple prétexte que la loi exige l'utilisation de "chiffres de recensement".

Les arguments sont différents en ce qui concerne la répartition du pouvoir politique, bien qu'à l'encore, l'obsession pour l'exactitude des chiffres du recensement à une période donnée n'ait pas sa raison d'être. Il est vrai que les chiffres du recensement servent aussi à la répartition des sièges à la Chambre des communes au Canada (et à la Chambre des représentants aux États-Unis). Cependant, cette répartition vaut pour dix ans. Pendant cette période, il se produit d'énormes déplacements de population. Si on laisse de côté l'interprétation des lois, il me semble que la question fondamentale est de savoir si, en redressant les chiffres du recensement en fonction du sous-dénombrement estimé, on peut réduire réellement une fonction de perte appropriée qui aurait pour but d'indiquer dans quelle mesure, en moyenne, la norme de représentation n'est pas respectée **sur une période de dix ans**. Je n'ai pas fait le calcul. Cependant, il me semble que les déplacements de population qui surviennent au cours de dix années ont beaucoup plus d'ampleur que les taux de sous-dénombrement estimés. Il est donc permis de se demander si les redressements – aussi significatifs soient-ils – que l'on opère pour les années de recensement (et la redistribution des sièges qui s'ensuit) n'auraient pas moins d'importance que les écarts par rapport à la norme de représentation qui sont causés par la migration sur une période de dix ans. Comme cet usage particulier du recensement est stipulé dans la constitution, il est inutile de vouloir modifier la loi. Toutefois, il conviendrait d'engager, sur la base de données scientifiques, des discussions sur l'interprétation à donner à la constitution, en prenant en considération les deux grandes causes de distorsion de la représentation politique dans la période intercensitaire: les erreurs du recensement et les déplacements de population (surtout la migration).

COMMENTAIRES

IVAN P. FELLEGI¹

Freedman et Navidi exposent avec beaucoup de précision et de lucidité les considérations et les arguments qui ont été présentés sur la question du redressement des chiffres du recensement de 1980 aux États-Unis. Ces arguments portent essentiellement sur la justesse des chiffres et des proportions de population enregistrés le jour du recensement. Par ailleurs, on tient pour acquis que, quelle que soit la décision qui sera prise au sujet du redressement, compte tenu de l'importance accordée aux chiffres de population, le redressement touchera la base de données du recensement dans son entier et, par voie de conséquence, tous ses sous-produits. Au lieu d'analyser en détail les arguments des parties en cause (bien que je croie qu'on a pris la bonne décision en ce qui concerne le recensement de 1980), j'aimerais exposer, d'un point de vue canadien, quelques idées qui ouvrent au débat un cadre de référence plus grand.

1. Le recensement est beaucoup plus qu'un dénombrement

Depuis les premiers recensements de l'époque contemporaine, l'objectif n'est pas uniquement d'obtenir un compte exact de la population. Pourtant, les ouvrages de plus en plus nombreux qui traitent la question du redressement des chiffres du recensement tendent à évaluer les avantages relatifs des méthodes proposées uniquement en fonction de l'estimation du nombre total (ou de la proportion) de personnes vivant dans un ensemble de régions. Je comprends évidemment pourquoi il en est ainsi: a) le problème est déjà assez complexe et b) les chiffres de population (ou estimations démographiques) mettent en jeu de très grosses sommes d'argent et beaucoup de pouvoir politique.

Je reviendrai au point b). Quant au point a), je pense qu'il n'est pas justifié, scientifiquement, de redresser les chiffres du recensement par une quelconque méthode sans prendre en considération les effets d'une telle opération sur la multitude d'usages que peuvent avoir les données et des proportions de population, nous tenterions fort probablement (du moins au Canada) d'élaborer des méthodes totalement différentes pour réaliser cet objectif. Compte tenu de la multiplicité des objectifs que vise le recensement, le fait qu'il est difficile de trouver un modèle adéquat pour cette base de données multidimensionnelle riche en informations n'est pas une raison valable pour s'intéresser uniquement aux chiffres de population et appliquer indistinctement les conclusions à toute la base de données.

2. L'exactitude des chiffres de population à une période donnée n'est peut-être pas le critère approprié pour la répartition intersectorielle des fonds publics et de la représentation politique

On semble se préoccuper à outrance de l'exactitude des chiffres et des proportions de population enregistrés l'année du recensement. Evidemment, si on engage des frais pour un recensement, c'est pour faire un dénombrement périodique qui produit des données justes et comparables pour de petites régions et de petits groupes de population. Mais la préoccupation excessive, me semble-t-il, pour l'exactitude des chiffres du recensement s'explique par des considérations d'équité: les chiffres de population déterminent en partie la répartition de grosses sommes d'argent et du pouvoir politique. Examinons ces deux aspects tour à tour.

D'abord, en ce qui concerne la répartition des fonds publics, il est vrai qu'au Canada des sommes considérables sont transférées de l'administration fédérale vers les administrations provinciales selon des formules qui reposent largement sur les chiffres et les proportions de population. Cependant, deux points méritent d'être soulignés sous le rapport des redressements.

¹ Ivan P. Fellegi, Statisticien en chef du Canada, Canada, Statistique Canada, Ottawa (Ontario) Canada, KIA 0T6.

produire une meilleure façon de redresser les chiffres, une méthode qui ne souffre pas de tous les défauts que renferment les méthodes préconisées par EKT. Mais cela ne signifie pas que le redressement effectué à l'aide de ces méthodes imparfaites n'aurait pas constitué une amélioration par rapport aux chiffres non redressés et contenant beaucoup d'imperfections dont on disposait.

4. Redressement des chiffres du recensement de 1990

À divers endroits dans tout l'article, les auteurs font allusion à des questions comparables et à des impondérables reliés au redressement des chiffres du recensement de 1990. Je pense que le lecteur devrait faire une distinction nette entre les méthodes utilisées dans les analyses présentées dans le cadre de la poursuite de 1980 et celles qui ont été employées comme partie intégrante du recensement de 1990. Un bon nombre des problèmes rencontrés par les personnes qui ont tenté de préparer des chiffres redressés en 1980 ont manifestement été éliminés et la discussion sur le redressement des chiffres de 1990 est beaucoup mieux définie. De plus, contrairement à la situation en 1980, les principaux méthodologistes statistiques du Census Bureau et la Directrice elle-même, ont trouvé les méthodes de redressement utilisées en 1990 justifiables et ils ont recommandé de redresser les chiffres du recensement. La proposition des statisticiens a été rejetée par le Secrétaire au Commerce. La question est maintenant, encore une fois, devant les tribunaux.

Freedman et Navidi n'énouencent pas leur position sur l'utilisation de techniques de redressement pour les chiffres du recensement de 1990, mais Freedman (1991) indique clairement que son opinion sur la situation de 1980 n'a pas changé. Je ne suis pas de cet avis. Il se peut bien qu'on ait raison de soutenir, comme le font les auteurs, que le Census Bureau n'aurait pas dû redresser les chiffres du recensement de 1980. Mais le recensement de 1990 est une autre question. En juin, dans un rapport du General Accounting Office, un organisme d'enquête du Congrès des E.-U., on a déclaré qu'il y avait 25.4 millions d'erreurs grossières dans le recensement de 1990, ce qui correspond à environ 10.4% de la population résidente. Le Bureau estime que le sous-dénombrement net était d'environ 5 millions de personnes et que la différence du taux de sous-dénombrement était la plus élevée depuis que le Bureau a commencé à estimer cette différence à partir du recensement de 1940. La méthodologie nécessaire pour effectuer le redressement des chiffres du recensement de 1990 est très améliorée par rapport à celle qui était au cœur du litige en 1980. Selon moi, les résultats des études d'évaluation du Census Bureau favorisaient nettement le redressement des chiffres du recensement de 1990. Il se peut que le juge qui est saisi du litige actuel ne voie pas la question du redressement de la façon dont Freedman et Navidi ont tendance à la formuler.

3. Questions pour lesquelles il y a désaccord avec EKT

Freedman et Navidi consacrent beaucoup de temps à ressasser la question de la multiplicité des séries du PEP et à faire ressortir les variations qui existent entre elles. Bien que la position selon laquelle il n'y a pas de choix clair et dominant parmi les différentes méthodes de redressement ait une certaine valeur, il se pourrait quand même que plusieurs des choix possibles donnent des résultats supérieurs à un recensement dont les chiffres ne sont pas redressés. Les auteurs se concentrent sur la variation parmi l'ensemble complet de douze possibilités, dont je trouve certaines peu plausibles compte tenu des hypothèses sur lesquelles ils se basent. Bien que je considère raisonnables les arguments invoqués pour appuyer l'utilisation de redressements synthétiques, je reconnais, avec les auteurs, qu'il existe une nette différence entre les redressements synthétiques et les redressements des estimations du PEP.

Où en sommes-nous dans ce débat? Je trouve la conclusion d'Ericksen, Kadane et Tukey irrésistible, bien que je sois d'accord avec Freedman et Navidi qu'il reste des questions à propos du choix précis des techniques favorisées par EKT. Freedman et Navidi soutiennent que la prétention principale d'EKT n'a aucun rapport avec la question de l'exacitude. Je ne suis pas de cet avis. Il se peut que les auteurs croient que les millions de personnes non dénombrées qui, pour presque tous les observateurs, ont été manquées en 1980 se cachent encore dans les contreforts des montagnes du Dakota du Sud ou dans un autre Etat avec peu de minorités.

Le cas des problèmes soulevés quand les hypothèses ne sont pas respectées constitue un thème familier dans divers écrits d'un des présents auteurs. Ici encore, les auteurs poursuivent ce thème relativement à l'équation linéaire employée pour effectuer le lissage. Ils semblent soutenir qu'il faut que toutes les hypothèses soient parfaitement justifiées, sinon "plus rien ne tient". Rien ne saurait être plus faux. Ils ne s'attendent certainement pas à ce qu'une personne quelconque accepte l'argument voulant que l'éruption du Mont St. Helens ait nui considérablement à la réalisation du recensement et que par conséquent cela réduit l'utilité de la méthode de lissage.

De même, je ne suis pas d'accord avec leur idée que le fait de préciser avec exactitude les variables explicatives est critique pour l'exacitude du lissage. Finalement, j'ai lu le rapport de Vivisaker (1991) qui a réexaminé les données provenant du recensement d'essai réalisé à Los Angeles en prévision du recensement de 1990, mais je n'ai pu y trouver la preuve qui, selon Freedman et Navidi, appuie leur prétention que le lissage augmente la variabilité.

Je crois, à l'instar de Freedman et Navidi, que le processus du recensement est extrêmement complexe et que la méthode de redressement qui a été proposée dans le cadre du litige portant sur le recensement de 1980 est loin d'être parfaite. Cependant, je trouve leurs arguments exagérés et je pense qu'ils ont tendance à perdre de vue la vieille maxime qui dit que "le mieux est l'ennemi du bien". Bien entendu, les hypothèses ne sont pas respectées. Bien entendu, on pourrait

L'exactitude des chiffres du recensement ainsi que le processus de redressement sont tous deux mis en cause. Et, c'est la *différence* importante du taux de sous-dénombrement, c.-à-d. la différence entre le sous-dénombrement des Noirs et le sous-dénombrement des personnes noires et entre celui qui touche les personnes d'origine hispanique et le sous-dénombrement relatif aux personnes d'origine non hispanique qui est importante quand nous en venons à évaluer l'exactitude des chiffres du recensement. Cela est dû au fait que les chiffres du recensement sont généralement utilisés pour répartir des ressources entre des groupes de la population, ressources comme les sièges à la Chambre des représentants des États-Unis; les sièges dans les législatures des États; les fonds versés par l'administration fédérale et ainsi de suite.

À l'aide de la méthode de l'analyse démographique, le Census Bureau a établi que, de 1940 à 1980, la différence dans le taux de sous-dénombrement des Noirs et des personnes non noires est demeurée à peu près constante, s'établissant entre 5% et 6%, bien que le sous-dénombrement global ait diminué de 5,6% à 1,4% (voir Fay et coll., 1988). Le nombre 1,4% ne signifie pas que le recensement a dénombré correctement plus de 98% de la population américaine en 1980. Ce nombre représente plutôt le sous-dénombrement *net*, qui peut être considéré comme la différence entre le sous-dénombrement réel (composé des personnes manquées ou des omissions) et le surdénombrement (personnes dénombrées par erreur et personnes dénombrées plus d'une fois). Même si les erreurs dues au surdénombrement et au sous-dénombrement se contrebalançaient parfaitement au niveau national, produisant un taux national de sous-dénombrement de 0%, il se pourrait que nous ayons encore un problème dû à la différence du taux de sous-dénombrement. Pour le recensement de 1980, le Bureau a déterminé que 6 millions de personnes avaient été dénombrées par erreur, dont jusqu'à un million étaient des inventions et jusqu'à 2,5 millions correspondaient à des personnes qui, par erreur, avaient été dénombrées deux fois au même endroit. Étant donné que le Bureau a déclaré un sous-dénombrement net de 1,4% ou de 3,2 millions de personnes en 1980, nous avons une estimation de 9,2 millions d'omissions (personnes qui ont été manquées) du chiffre total du recensement de 1980. Si l'on additionne les omissions au nombre de personnes dénombrées par erreur, nous obtenons un total de 15,2 millions d'erreurs dans le dénombrement des personnes, ce qui correspond à presque 7% du total officiel du recensement de 1980. Selon moi, ce niveau d'erreur pour le recensement représente un problème important qu'on doit aborder quand nous parlons de l'opportunité de redresser les chiffres du recensement de 1980. Bien entendu, peu après la fin du recensement de 1980, le Census Bureau a dressé un tableau beaucoup plus rose de l'exactitude des chiffres bruts du recensement. Il se peut que, conformément à la signification littérale du titre de l'article, Freedman et Navidi espèrent que nous accepterons comme exact ce que nous savons maintenant avoir été une évaluation grossièrement incomplète de la part du Census Bureau. J'espère dans ce n'est pas le cas. Nous en savons maintenant beaucoup plus sur le niveau d'erreur dans les chiffres bruts du recensement de 1980. La question restante est de savoir si nous avons de meilleures informations sur les diverses formes de chiffres redressés étant donné qu'une décennie s'est écoulée.

2. Faits et théorèmes

L'article dont je fais la critique est plein d'énoncés à propos de l'exactitude des procédures de redressement des chiffres du recensement. Quand il s'agit d'énoncer et de prouver des théorèmes, je ne doute pas que Freedman et Navidi le fassent correctement. La pertinence de tels théorèmes pour le redressement des chiffres du recensement est une autre question. Freedman et Navidi présentent un contre-exemple simple et apparemment irrésistible au théorème de Schirm-Preston sur le redressement synthétique. Il est certainement vrai que les totaux globaux pour l'État A et pour l'État B dans leur exemple sont exacts pour le recensement et inexacts (mais à peine) au niveau du redressement synthétique. Mais il est aussi vrai que le changement considérable dans le nombre de Blancs et de Noirs dans l'État B est ce que, selon moi, un redressement est conçu pour effectuer et que la correction est réalisée au prix d'une

COMMENTAIRES

STEPHEN E. FIENBERG¹

Freedman et Navidi livrent leur dernière analyse favorisant la réflexion sur la question du sous-dénombrement lors du recensement décennal des États-Unis de 1980. Malheureusement, ils ne s'attaquent pas à la question posée dans le titre de l'article et tentent plutôt de justifier les opinions qu'ils avaient déjà exprimées dans Freedman et Navidi (1986) et de réfuter les commentaires faits par d'autres personnes sur les opinions présentées dans ce dernier article. Leur thème est familier à ceux qui ont lu des versions antérieures de la discussion portant sur le "procès de 1980" sur le redressement: le recensement est très complexe et on relève toujours un niveau, modeste, de sous-dénombrement; le redressement fait appel à la modélisation statistique qui est basée sur des hypothèses non vérifiables; un mauvais redressement peut être pire que pas de redressement du tout.

Je ne suis pas d'accord avec un bon nombre des opinions exprimées par les auteurs et je pense qu'ils déforment ce qui aurait dû être en jeu relativement à la correction des données du recensement de 1990. Dans le texte qui suit, je tente d'expliquer mes divergences d'opinions avec les auteurs et je donne mon point de vue sur deux questions: celle qui est soulevée dans le titre de l'article de Freedman et Navidi et celle qui est exprimée implicitement dans les données qui y sont présentées relativement au recensement de 1990. (Notez: L'auteur n'a joué aucun rôle dans le litige portant sur le redressement des chiffres du recensement de 1980, mais il apporte son concours à la ville de New York et à d'autres plaignants en ce qui concerne la poursuite découlant de la décision du Département du Commerce de ne pas redresser les résultats du recensement de 1990.)

1. Le titre et l'article portent sur deux questions différentes

Aurions-nous dû redresser les chiffres du recensement de 1980? Selon moi, la seule façon raisonnable de répondre à cette question est de se la poser dans le contexte des preuves disponibles à ce moment, ou du moins en tenant compte des preuves qui étaient disponibles quand les tribunaux se prononçaient sur la question. En soi, la description des questions définies par le Bureau of the Census et présentées dans la première partie de l'article est importante, bien que ces questions aient eu peu à voir avec la décision originale de ne pas redresser les chiffres du recensement de 1980, qui fut prise par le Directeur du Bureau avant que des renseignements sur la couverture ne soient disponibles.

Toutefois, le reste de l'article ne porte pas sur cette question. Il s'attaque plutôt à la tentative répétée par les défenseurs des deux parties de rassembler des preuves afin d'appuyer les positions qu'ils ont prises lors du litige. Essentiellement, les auteurs posent une question à propos des preuves dont on dispose actuellement pour appuyer une décision prise il y a une décennie. Comme pour toutes les questions statistiques, la poursuite de l'analyse des données et l'examen rétrospectif peuvent mettre à jour notre jugement sur la réponse à une telle question et on ne peut donc qu'approuver l'effort des auteurs pour reprendre, encore une fois, la preuve reléguée au recensement de 1980.

Nous pouvons donc passer à la formulation de la question à laquelle on doit répondre. Selon moi, le juge n'a pas exposé correctement les questions faisant l'objet du litige, comme c'est aussi le cas pour Freedman et Navidi quand ils décrivent la question du sous-dénombrement. Ils laissent entendre que la seule véritable question porte sur l'exactitude du processus de redressement et qu'il n'y a qu'un sous-dénombrement potentiellement petit à propos duquel nous devrions nous inquiéter. Aucune de ces deux positions ne pourrait être plus loin de la vérité.

¹ Stephen E. Fienberg, York University, North York (Ontario) Canada M3J 1P3.

SCHIRM, A.L., et PRESTON, J. (1987). Census undercount adjustment and the quality of geographic population distributions (avec discussion). *Journal of the American Statistical Association*, 82, 965-990.

SCHIRM, A.L. (1991). The effects of census undercount adjustment on congressional apportionment. *Journal of the American Statistical Association*, 86, 526-541.

WOLTER, K. (1986). Comment. *Statistical Science*, 1, 24-28.

WOLTER, K. (1991). Accounting for America's uncounted and miscounted. *Science*, 253, 12-15.

WOLTER, K., et CAUSEY, B. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

YLIVISAKER, D. (1991). A look back at TARO. Rapport technique, Department of Mathematics, UCLA.

Pennsylvanie et une perte pour la Californie. Cette mauvaise répartition des sièges entraînera peut-être une perte sociale nette mais il paraît tout à fait simpliste de vouloir mesurer cette perte au moyen de l'expression (16) ou de toute autre formule semblable.

Pretons un autre exemple. Pour simplifier le problème, supposons que le sous-dénombrement sous-estime la proportion de la population qui demeure dans ces villes et un redressement corrigera en partie la situation.

Or, comme le redressement se fait à l'aide de la méthode synthétique, les proportions de population seront modifiées pour toutes les régions. Dans les régions qui comptent une forte proportion de Noirs et de personnes d'origine hispanique, les parts de population seront haussées artificiellement au détriment d'autres régions. Ce redressement touchera même les régions où les résultats du recensement sont exacts.

Dans cet exemple, le redressement peut contribuer à une répartition plus juste des ressources entre les quatre grandes villes et d'autres régions mais il créera des distorsions un peu partout ailleurs. Dans ce genre de circonstances, la méthode de la fonction de perte est inefficace. L'équilibrage des inégalités est un problème politique, que l'on peut difficilement résoudre par une formule statistique.

Certains observateurs diront que cet exemple est exagéré. Or, seulement 5,000 îlots sont échantillonnés dans l'Enquête post-censitaire et le redressement doit s'appliquer à 39,000 divisions administratives de divers niveaux. Les données véritables sur le sous-dénombrement se limitent nécessairement à un petit nombre de localités. Pour ce qui est des autres régions, le redressement doit reposer surtout sur de la théorie et non des données.

BIBLIOGRAPHIE

CITRO, C.F., et COHEN M.L. (Éds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C. National Academy Press.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., et TURKEY, J.W. (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927-943.

ERICKSEN, E.P., ESTRADA, L.F., TURKEY, J.W., et WOLTER, K.M. (1991). Report on the 1990 Decennial Census and the Post Enumeration Survey, soumis au Secretary of the Department of Commerce, 22 juin 1991.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., et COWAN, C.D. (1988). *The Coverage of the Population in the 1980 Census*. Washington, D.C.: U.S. Department of Commerce, Government Printing Office.

FELLEGI, I. (1985). Comment. *Journal of the American Statistical Association*, 80, 116-119.

FREEDMAN, D.A. (1991). Adjusting the 1990 census. *Science*, 252, 1233-1236. Copyright 1991 par l'AAAS. Cité avec l'autorisation.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (avec discussion). *Statistical Science*, 1, 1-39.

HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.

ISAKI, C., DIFFENDAH, G., et SCHULTZ, L. (1987). Report on statistical synthetic estimation for small areas. Rapport technique, Bureau of the Census.

PASSEL, J. (1987). A note about synthetic estimates of undercount. Note de service, U.S. Bureau of the Census.

Voilà pour la méthode de redressement proposée. Nous revenons maintenant aux points b), c) et d).

b) Les estimations synthétiques sont peu efficaces lorsqu'il y a aggrégation. Fellegi (1985) l'a déjà souligné. Voir Cohen et Citro (1985, p. 318). Pour un autre exemple, voir les tableaux 3 à 5 ci-dessus.

c) Au niveau de l'îlot, l'erreur d'arrondissement peut être prépondérante. Le redressement des chiffres du recensement doit se faire effectivement au niveau de l'îlot. (L'îlot est la plus petite unité de la géographie du recensement; on compte 6,5 millions d'îlots au pays.) On peut trouver jusqu'à 25 "strates *a posteriori*" dans un îlot typique de région urbaine; à chaque combinaison "îlot-strate" correspond seulement un petit nombre de personnes. Donc, lorsqu'on multiplie l'effectif par un facteur de correction, on se trouve à ajouter ou à soustraire un nombre infime de personnes et les nombres fractionnaires sont normalement arrondis. L'exemple suivant montre bien comment l'erreur d'arrondissement peut effacer les avantages du redressement synthétique.

Supposons qu'il faille redresser l'effectif de n "régions"; il peut s'agir d'îlots combinés chacun avec une "strate *a posteriori*" déterminée. Supposons aussi que l'effectif recensé de chacune de ces régions est le même, c . Soit $m < n$. Supposons que dans chacune de m régions, une personne n'a pas été recensée; dans les $n - m$ autres régions, le chiffre du recensement est exact. En tout, m personnes n'ont pas été dénombrées. Tous ces renseignements sont connus, mais nous ignorons dans quels îlots se trouvent les personnes oubliées. Selon (16),

perte découlant de l'utilisation des chiffres non redressés du recensement = m/c . (18)

Le redressement s'effectuerait de la manière suivante: choisir m régions au hasard et ajouter une personne à l'effectif de chacune de ces régions. De toute évidence, perte prévue découlant du redressement

$$\frac{m}{n} \cdot m \cdot 0 + \left(1 - \frac{n}{m}\right) \cdot m \cdot \frac{1}{m} + \frac{n}{m} \cdot (n - m) \cdot \frac{1}{m} + \left(1 - \frac{n}{m}\right) \cdot (n - m) \cdot 0 = 2 \left(1 - \frac{n}{m}\right) \cdot \frac{n}{m} \cdot \frac{c}{m}. \quad (19)$$

Lemme. Si $m < n/2$, une perte nette découlera du redressement synthétique.

Démonstration. Si $m < n/2$, alors

$$2 \left(1 - \frac{n}{m}\right) \cdot \frac{n}{m} \cdot \frac{c}{m} > \frac{c}{m}. \quad (20)$$

Evidemment, cet exemple est presque aussi simpliste que le lemme (15). Bref, on ne peut établir la valeur d'une méthode de redressement par des arguments *a priori*.

d) Les fonctions de perte ne reproduisent qu'en partie le problème de politique et peuvent être plus nuisibles qu'utiles. Prenons tout d'abord un exemple. Supposons qu'il y ait des erreurs dans le recensement et que la principale conséquence de ces erreurs soit de faire perdre un siège du Congrès à la Californie au profit de la Pennsylvanie. Il y a donc un gain pour la

Alors,

(16)

$$\sum_n^{t=1} (x_t - t_t)^2 / c_t$$

est minimisée lorsque

$$\lambda = \left[\sum_n^{t=1} t_t \right] / \left[\sum_n^{t=1} c_t \right]$$

La démonstration est omise car évidente. La "fonction de perte" définie en (16) diffère à plusieurs égards de celle utilisée en (7-8-9) et en (13-14) et qui peut être exprimée par la formule

(17)

$$\sum_n^{t=1} \frac{C}{c_t} \left[\frac{X}{x_t} - \frac{T}{t_t} \right]^2,$$

ou

La fonction de perte (17) met l'accent sur les proportions tandis que la fonction (16) met l'accent sur les chiffres de population; de plus, (17) accorde plus d'importance aux grandes sous-populations alors que pour (16), c'est le contraire, à cause de la division par c_t . Nous n'attachons pas d'importance particulière à la fonction (17) et nous ne voyons pas comment nous pourrions préférer l'une ou l'autre fonction.

Le lemme (15) est mathématiquement exact mais il est tellement éloigné de la réalité du recensement de 1990 qu'il paraît virtuellement hors de propos. À ce sujet, il y a quatre points à considérer:

- a) L'effectif réel de la population totale, T , est inconnu; Wolter et Causey se penchent sur le problème, mais l'exemple du tableau 11 repousse leur argument; le redressement synthétique diminue l'exactitude des chiffres du recensement de 1980.
- b) Les estimations synthétiques sont peu efficaces lorsqu'il y a aggrégation.
- c) Au niveau de l'ilot, l'erreur d'arrondissement peut être prépondérante.
- d) Les fonctions de perte ne reproduisent qu'en partie le problème de politique et peuvent être plus nuisibles qu'utiles.

Nous allons traiter plus en détail les points b), c) et d) mais avant, jetons un coup d'oeil sur les méthodes qui ont été proposées pour redresser les chiffres du recensement de 1990. La population est divisée en 1,392 "strates formées *a posteriori*" (par ex.: locataires de sexe masculin et d'origine hispanique âgés de 30 à 44 ans et vivant dans les villes centrales de la division du Pacifique). Désignons ces strates par $f = 1, \dots, 1,392$. Pour chaque "strate *a posteriori*" f , on calcule un facteur de correction λ_f par des techniques de saisie-resaisie qui utilisent des données provenant d'une enquête post-censitaire (Freedman 1991).

Les 1,392 facteurs servent à redresser les chiffres de population pour chaque petite région. Voici comment. Choisissons une région, par exemple une petite ville. On trouvera un grand nombre de strates formées *a posteriori* dans cette région. On multiplie l'effectif recensé pour chaque combinaison "région-strate" par le facteur λ_f correspondant et on fait la somme des produits. Autrement dit, on redresse l'effectif de sous-populations à l'aide de la méthode synthétique, puis on additionne les estimations synthétiques afin d'obtenir des totaux pour les petites régions.

Tableau 11

Proportions de population établies d'après les chiffres du recensement, les estimations synthétiques B et la série d'estimations PEP 2-9, en pourcentage; chiffres du recensement: en milliers.

	Nord-Est	Midwest	Sud	Ouest	Total
Synthétiques B - PEP 2-9	.08%	.03%	.24%	-.35%	.00%
Recensement - PEP 2-9	.10%	.06%	.12%	-.28%	.00%
PEP 2-9	21.59%	25.92%	33.15%	19.34%	100.00%
Synthétiques B	21.67%	25.95%	33.39%	18.99%	100.00%
Recensement	21.69%	25.98%	33.27%	19.06%	100.00%
Population recensée	49,135	58,866	75,372	43,172	226,545

Le tableau 11 contient les données pertinentes pour les quatre régions de recensement: Nord-Est, Midwest, Sud et Ouest. Notons aussi les carrés des écarts entre les parts de population, pondérées par la taille:

- (13) é.q.m. entre les estimations synthétiques B et la série PEP 2-9 = 0.21 de 1%.
- (14) é.q.m. entre les chiffres du recensement et la série PEP 2-9 = 0.15 de 1%.

La série PEP 2-9 est assez comparable à la "moyenne des séries du PEP jugées supérieures" du tableau 2. Dans ce tableau-là, les chiffres du recensement se rapprochaient plus des estimations synthétiques B que des estimations du PEP. Dans le tableau 11, les chiffres du recensement se rapprochent plus des estimations du PEP, et les estimations synthétiques sont les valeurs aberrantes. Le mode de décomposition semble faire la différence entre les deux tableaux. Dans le tableau 2, la population est décomposée selon l'origine raciale et ethnique tandis que dans le tableau 11, elle l'est selon la géographie classique du recensement. Evidemment, si nous utilisons un autre mode de décomposition ou une autre méthode de redressement synthétique, nous pourrions obtenir des résultats contraires à ceux observés dans le dernier tableau; une modification de la fonction de perte pourrait avoir les mêmes conséquences. Afin d'illustrer les possibilités, envisageons de redresser les chiffres de population des 66 régions visées par le PEP plutôt que ceux des quatre régions de recensement. Considérons toujours la série PEP 2-9 comme la "réalité". Avec la fonction de perte (17), les chiffres du recensement surpassent de peu les estimations synthétiques B. Avec la fonction (16), les estimations synthétiques B accusent une perte beaucoup moindre que les chiffres du recensement.

Fonctions de perte

Les partisans du redressement dans la cause du recensement de 1990 fondent leurs arguments analytiques sur les fonctions de perte; voir Wolter et Causey (1991) ou Ericksen, Estrada, Tukey et Wolter (1991, p. 20 du rapport principal; annexes C et H). L'essentiel de leur raisonnement peut se résumer dans le lemme ci-dessous. Quelques mots sur la notation: le pays est divisé en n régions désignées par i ; c_i est la population recensée dans la région i et t_i la population réelle. L'"estimation synthétique" pour la région i est $x_i = \lambda c_i$, où le "facteur de correction" λ est calculé à l'aide d'autres données.

Lemme. Pour $i = 1, \dots, n$ posons $c_i > 0$ et $t_i > 0$. Soit $0 < \lambda < \infty$ et $x_i = \lambda c_i$. (15)

9. QUELLE A ÉTÉ LA DÉCISION DU TRIBUNAL?

Au moment de la rédaction de cet article, la question du recensement de 1990 était toujours en litige. Pour ce qui est du recensement de 1980 toutefois, le tribunal a statué en faveur du défendeur sur tous les points. Voici un extrait de la décision *Cuomo et al. v. Baldirige et al.* 674 F. Supp. 1089-1108 (SDNY 1987).

“ [...] Les autorités de l'Etat et de la ville [de New York] ont intenté une action contre le secrétaire au Commerce, le directeur du Bureau of the Census et d'autres représentants fédéraux dans le but de faire redresser les chiffres du recensement décennal de 1980. Le juge Sprizzo, de la cour de district, a conclu que les représentants de l'Etat et de la ville n'étaient pas parvenus à prouver que le redressement statistique des chiffres du recensement décennal était techniquement faisable.” (TRANSDUCTION)

“ [...] Pour qu'une telle opération puisse se faire, il faut pouvoir compter sur une méthode de redressement qui donnera une image plus juste de la population des Etats-Unis dans chaque Etat, aux fins de la répartition des sièges au Congrès, et dans chaque partie d'Etat, aux fins de la répartition des fonds publics. [...] Si ce n'est pas le cas, il faut écarter tout projet de redressement [...] car [...] les sièges au Congrès aussi bien que les fonds publics sont des quantités fixes et un accroissement de la population dans un Etat ou une partie d'Etat signifiera nécessairement une diminution de la part destinée à d'autres régions [...] (TRANSDUCTION)

“Malgré la complexité des faits [...] la cour doit se prononcer sur une question en particulier, à savoir si les demandeurs se sont acquittés de la charge de prouver que des chiffres redressés reflètent mieux la répartition géographique réelle de la population des Etats-Unis que des chiffres non redressés. La cour conclut d'après les faits que les demandeurs ne se sont pas acquittés de cette charge et en conséquence, l'action doit être rejetée [...]” (TRANSDUCTION)

REMERCIEMENTS

Freedman étudie actuellement pour le compte du Département de la justice des questions qui découlent du recensement de 1990. Cependant, le département ne partage pas nécessairement les opinions exprimées dans cet article. Les auteurs remercient L. Bazel (San Francisco), P. Diaconis (Harvard), S. Klein (RAND) et A. Tversky (Stanford) pour leurs commentaires utiles.

ANNEXE

Estimation synthétique et fonctions de perte

Estimation synthétique

Dans la section 5 de Wolter et Causey (1991), on démontre empiriquement que le redressement synthétique aurait amélioré les chiffres du recensement de 1980. Pour cela, on se sert d'une étude de simulation où "recensement" et "réalité" sont définis tous deux en fonction d'une population de référence artificielle élaborée par Isaki et coll. (1987). Toutefois, comme le montre Passel (1987), l'argument dépend assez largement de la population de référence. Notre but, ici, est de reprendre un des exemples de Passel en y apportant des modifications. En effet, si on définit la population de référence en se servant de la série PEP 2-9 pour redresser les chiffres du recensement de 1980, le redressement synthétique donne des résultats qui s'éloignent de la réalité.

Nous avons fait une autre simulation en considérant cette fois la série FEP 2-9 comme la "réalité" et en incluant la "proportion de population urbaine" parmi les variables explicatives qui peuvent être choisies. Les résultats figurent dans le tableau 10. Là encore, la "proportion de la population recensée selon la méthode classique" revient fréquemment, tout comme le "pourcentage de minorités". Dans le cas des autres variables, les résultats sont plutôt incohérents. Et la très vilaine "proportion de population urbaine" est choisie plus souvent que cinq des variables d'EKT, y compris l'indicateur de ville centrale. Non, les données ne déterminent pas le modèle.

7.4 Modèle de régression

En tant que statisticiens, nous sommes intéressés par des arguments sur la régression. Cependant, le tribunal n'a pas été impressionné:

"Dans leur réplique, les demandeurs ont prétendu qu'en soumettant les estimations du sous-dénombrement tirées du PEP à l'analyse de régression, le Bureau pourrait par la suite se servir du PEP pour redresser avec précision les chiffres du recensement de 1980. Or, les experts des deux parties sont d'accord pour dire que l'analyse de régression ne peut d'aucune manière atténuer le biais contenu dans le PEP et les demandeurs ne semblent pas prétendre le contraire. Bref, bien que l'analyse de régression puisse réduire l'erreur d'échantillonnage aléatoire dans le PEP, elle ne réduira pas les erreurs substantielles causées par de mauvais appariements, par les hypothèses non vérifiées relatives aux cas non résolus et par le biais de corrélation. De plus, la preuve très étoffée présentée devant le tribunal appuie les conclusions des experts du défendeur, selon lesquelles les principales difficultés du PEP tiennent plus à ces biais qu'à l'erreur d'échantillonnage." (TRADUCTION) (674 F Supp 1103, citations et renvois omis.)

8. RÉSUMÉ ET CONCLUSION

Ericksen, Kadane et Tukey prétendent qu'ils peuvent améliorer les chiffres du recensement de 1980 par des opérations de redressement. Ils semblent maintenant reconnaître qu'il aurait été injustifié d'effectuer un redressement pour les sous-régions des 66 régions visées par le PEP. En ce qui a trait aux régions proprement dites, le différend demeure. À notre avis, l'efficacité de l'un ou l'autre des redressements proposés par EKT repose sur des hypothèses non vérifiées et peu plausibles concernant les données manquantes, les mécanismes du sous-dénombrement, le biais contenu dans le PEP et les erreurs stochastiques dans les modèles de régression. Un changement d'hypothèses entraîne une modification des résultats et, à notre humble avis, le fait de calculer des moyennes par rapport à diverses séries d'hypothèses n'élimine pas le problème. EKT concluent en ces termes (p. 943):

"Nous croyons que le Census Bureau s'attire des ennuis politiques en ne faisant aucun cas du sous-dénombrement. Il gagnerait à mettre tout en oeuvre pour redresser les chiffres du recensement à l'aide des méthodes statistiques et démographiques dont il dispose. Des erreurs subsisteront mais elles seront moindres, et nous ne saurons plus d'avance qui perd de l'argent et de la représentation à cause du sous-dénombrement." (TRADUCTION)

Cette analyse politique n'est pas sans fondement, sauf qu'une mise au point s'impose. Nous trouvons tout à fait injuste de dire que le Census Bureau ne fait aucun cas du sous-dénombrement. De même, les ennuis politiques du Bureau ne sont pas le fait de sa seule conduite. On peut en effet imaginer des redressements pour satisfaire des groupes particuliers ou régler des actions judiciaires. Or, les résultats du recensement servent de base à la répartition d'une quantité fixe de ressources, de sorte qu'il y aura toujours des gagnants et des perdants. Ces gagnants et ces perdants n'auront aucune difficulté à se reconnaître par la suite, si ce n'est dès le départ. De plus, l'objectif d'accroître l'exactitude des chiffres du recensement au moyen d'un redressement statistique s'est révélé jusqu'ici illusoire.

Simulation pour le choix de variables; série PEP10-8 considérée comme la "réalité".

Tableau 9

Essai	VC	Min	Crim	Class	Ed	Pauv	Lang	LM
1			x	x		x		
2		x				x		
3		x				x		
4	x		x			x		
5	x					x		
6		x				x		
7				x		x		
8			x			x		
9	x					x		
10	Aucun modèle ne répondait aux critères d'EKT							

Notes: VC est l'indicateur de ville centrale; Min, le pourcentage de minorités; Crim, le taux de criminalité; Class, la proportion de la population qui a été recensée selon la méthode classique; Ed, la proportion de personnes qui n'ont pas de diplôme d'études secondaires; Pauv, le pourcentage de la population qui vit sous le seuil de la pauvreté; Lang, la proportion de personnes qui maîtrisent difficilement la langue anglaise; LM, le pourcentage de la population qui habite un immeuble à logements multiples.

Tableau 10

Simulation pour le choix de variables. Série PEP 2-9 considérée comme la réalité; la "proportion de population urbaine" (Urb) est admise parmi les variables explicatives. Le tableau indique le nombre de fois qu'a été choisie chaque variable et la valeur moyenne du coefficient correspondant (pour le nombre de fois qu'a été choisie la variable); 100 séries de données ont été produites.

Variable	Nombre de fois sélectionnée	Valeur moyenne du coefficient
VC	17	2.954
Min	82	0.071
Crim	53	0.053
Class	93	0.028
Ed	5	0.085
Pauv	1	0.135
Lang	17	0.315
LM	0	*****
Urb	23	0.060

Faisons ici une courte digression sur les méthodes du recensement. La "méthode de recensement classique" consiste à demander aux répondants de remplir un questionnaire de recensement et de le remettre en mains propres à un recenseur lorsque celui-ci se présentera à leur domicile; cette méthode est appliquée dans les régions en grande partie rurales, particulièrement dans l'Ouest. Class désigne le pourcentage de la population qui vit dans des régions ayant fait l'objet d'un recensement classique. (Dans les régions urbaines, il fallait retourner le questionnaire par la poste.) En 1980, le taux de sous-dénombrement était relativement élevé dans les régions rurales à cause probablement de cartes et de listes d'adresses incomplètes, ce qui explique peut-être pourquoi class est une variable explicative aussi forte.

semblent tout aussi erronées. Bien sûr, les régions urbaines sont différentes entre elles, comme le disent si bien ces auteurs. Il en est de même des villes centrales. De la même manière, les membres de minorités qui vivent dans les villes centrales ont des chances d'être différents de ceux qui vivent dans les banlieues. Et ainsi de suite. Toutes les variables d'EKT sont des "prédicteurs vagues" du sous-dénombrement et certaines d'entre elles sont plus vagues que la "proportion de population urbaine" (p. 934).

Sur cet aspect du litige, le juge est plus sévère que nous à l'égard d'Ericksen et Kadane: "[...] De plus, comme l'ont si bien exposé les témoins experts du défendeur, on ne peut démontrer de façon sûre qu'une des séries d'estimations du PEP est supérieure aux autres ou même, au recensement proprement dit car les renseignements dont on dispose ne permettent pas de déterminer quels procédés du PEP conviennent le mieux pour mesurer le sous-dénombrement dans le recensement. Tandis que deux des témoins experts du demandeur ont exprimé leur préférence pour les estimations de la série "PEP-2-9" en se fondant sur l'hypothèse que les procédés du PEP qui ont permis d'obtenir ces estimations sont supérieurs aux procédés qui ont servi à produire les autres estimations du PEP, ces experts n'ont pu étayer leur affirmation que par des hypothèses non vérifiées. Par contre, les témoins experts du défendeur ont présenté des hypothèses tout aussi plausibles qui tiennent pour supérieurs d'autres procédés du PEP qui donnent des résultats totalement différents." (TRADUCTION) (674 F Supp 1102, citations et renvois omis.)

Nous avons fait une étude de simulation qui visait à démontrer trois choses: a) les données ne peuvent servir à déterminer quelles variables entreraient dans le modèle; b) les erreurs-types dépendent des hypothèses concernant les termes d'écart et c) les erreurs-types calculées par Ericksen et Kadane étaient très optimistes. Nous voulions aussi vérifier deux autres points: d) les erreurs-types ne permettent pas de mesurer l'effet d'un biais; e) la technique de lissage d'Ericksen-Kadane ne fait aucun cas des biais du PEP qui sont étroitement liés aux variables explicatives.

Compte tenu des cinq points ci-dessus, il est difficile de démontrer que le modèle proposé améliore la qualité des estimations du PEP. EKT ne commentent pas les points b), d) et e). Ils réfutent a) mais concèdent plus ou moins le point c). Quant à nous, nous admettons que dans notre simulation, laquelle reconnaît la motilité du modèle, la régression réduit effectivement l'erreur d'échantillonnage. Nous maintenons toutefois que a) est juste, comme nous le verrons plus loin. Par ailleurs, dans d'autres circonstances, le lissage peut accroître l'erreur d'échantillonnage (Ylvisaker 1991, p. 7).

EKT (p. 943) critiquent notre étude de simulation parce qu'elle ne visait que les modèles à équation à trois variables et qu'elle ne restreignait en rien les valeurs de la statistique t . Nous allons donc répéter la simulation ici. Essentiellement, il s'agit de considérer la série PEP 10-8 comme la "réalité" et d'introduire, pour chacune des 66 régions étudiées i , une erreur aléatoire de variance K_i , conformément à (4). Nous admettons par le fait même l'équation (1) et les hypothèses relatives à δ_i . Nous choisissons les variables suivant la méthode décrite par EKT (p. 935), puis nous ajustons le modèle de régression; nous répétons l'opération 100 fois.

Le tableau 9 indique les variables qui ont été choisies dans les dix premiers essais. Les résultats sont variables sauf pour ce qui a trait à la "proportion de la population recensée selon la méthode classique", qui a été sélectionnée à chaque essai. Pour l'ensemble des 100 essais, à l'exclusion de ceux qui n'ont pas produit de modèle satisfaisant, l'é.q.m. nominal était d'environ 30% inférieur à la normale et la différence de qualité entre les estimations composites et les estimations du PEP, en faveur de celles-ci, était gonflée de 75%. C'est là que nous voyons toute l'importance des hypothèses.

Ecart quadratique moyen des équations de régression pour les séries PEP 2-9 et PEP 10-8. Les variables explicatives sont la proportion de minorités dans la région étudiée, la proportion de la population recensée selon la méthode classique, et soit le taux de criminalité ou la proportion de population urbaine.

Taux de criminalité		Proportion de population urbaine	
PEP 2-9	1.53	1.54	
PEP 10-8	1.35	1.33	

Tableau 8

ET (ε) et ET de l'é.q.m. pour les 66 régions étudiées; PEP 2-9 et PEP 10-8. Les modèles comprennent la proportion de minorités dans la région étudiée, la proportion de la population recensée selon la méthode classique, et soit le taux de criminalité ou la proportion de population urbaine.

Taux de criminalité		Proportion de population urbaine	
ET (ε)	ET de l'é.q.m.	ET (ε)	ET de l'é.q.m.
PEP 2-9	.75	.76	.65
PEP 10-8	.00	.00	.25

Notes: Soit K une matrice diagonale 66×66 dont l'élément (i,i) est K_i . Désignons par X la matrice 66×4 des variables explicatives. Soit $H = X(X^TX)^{-1}X^T$ et $I - H = K^{-1} + ET(\epsilon)^{-2}$ par l'expression $IK^{-1}y$, où y est le vecteur 66×1 des estimations du PEP. L'ET de l'é.q.m. pour les 66 régions étudiées est $\sqrt{\text{trace } I/66}$. Pour plus de détails, voir FN. À l'audience, Ericksen et Kadane avaient calculé ET (ε) pour 51 régions (les 50 Etats plus le District de Columbia); nous avons fait de même dans FN. Cette fois-ci, nous utilisons 66 régions, comme semblent le recommander EKT. La différence est notable.

Sur ces idées, retournons à la question initiale: quelle série du PEP choisir et laquelle de ces deux variables explicatives doit-on retenir: taux de criminalité ou proportion de population urbaine. Pour autant que nous puissions en juger, d'après les critères choisis par EKT, la différence entre le taux de criminalité et la proportion de population urbaine est insignifiante. Et la série PEP 10-8 est nettement supérieure à la série 2-9. Voir le tableau 7.

À la page 935 et 940 d'EKT, σ désigne l'écart quadratique moyen. Il y a une certaine confusion dans les symboles car on utilise σ^2 pour désigner $\text{Var}(\epsilon)$ dans les équations (2) et (4), suivant Ericksen et Kadane (1985, p. 105) ou FN (p. 5). Pour éviter toute confusion, désignons par ET (ε) la valeur estimée de σ , écart quadratique moyen; de cette valeur découle l'erreur-type de chacun des 66 taux de sous-dénombrement calculés à l'aide du modèle d'Ericksen-Kadane, comme l'indiquent les équations (8) et (10) dans FN. Pour ce qui a trait à la série PEP 10-8, la valeur estimée ET (ε) est pratiquement nulle, de sorte qu'un modèle fondé sur 10-8 donne un très bon ajustement et les 66 taux de sous-dénombrement sont estimés avec beaucoup de précision (tableau 8).

Pour ce qui a trait aux "critères statistiques", contrairement à ce qu'affirment EKT, la série 10-8 est supérieure à la série 2-9 et la proportion de population urbaine n'est pas une moins bonne variable explicative que le taux de criminalité. Sur le plan qualitatif, les remarques d'EKT

de point de comparaison (10-8). Les arguments ont été analysés devant le tribunal et dans FN (p. 8, l'analyse, et p. 36, la réponse à la réplique). Nous restons sur nos positions: il n'y a pas de raison formelle de choisir la série 2-9 plutôt que la 10-8.

EKT disent que nous avons prôné l'idée que la "proportion de population urbaine" devait être considérée comme une variable indépendante (p. 934). Ce n'est pas tout à fait exact. Nous estimions que le choix de variables indépendantes fait par EKT était quelque peu arbitraire et nous voulions montrer qu'un changement de variables pouvait modifier grandement les résultats - encore l'analyse de sensibilité. Des différences ont été notées surtout dans le cas des petites régions (FN, p. 9). Comme EKT ne préconisent plus le redressement des chiffres du recensement de 1980 pour les petites régions, cet argument n'a peut-être plus d'intérêt pratique.

Voyons maintenant un autre aspect de la question: EKT soutiennent qu'il faut choisir les modèles en se fondant sur des critères statistiques (p. 941). Essentiellement, ils recommandent de choisir les variables de manière à ce que l'é.q.m. d'un ajustement par les MCO soit minimum. Or, l'é.q.m. sert à mesurer le degré d'association des données et non la validité de la théorie qui est en question.

Pour des raisons qui demeurent obscures, EKT limitent leur analyse à des modèles à 2, 3 ou 4 variables et ils rejettent les coefficients qui ont une valeur *t* inférieure à 2. L'équation qu'ils préfèrent semble être la suivante:

$$\text{PEP 2-9} = -2.23 + .079 \text{ min} + .036 \text{ crime} + .028 \text{ class} + \text{résidu} \quad (11)$$

é.q.m. = 1.53.

(-4.0) (5.4) (3.6) (3.5)

Les variables de l'équation sont la proportion de minorités dans la région étudiée, le taux de criminalité et la proportion de la population recensée selon la méthode classique; les valeurs *t* figurent entre parenthèses; on calcule l'é.q.m. à l'aide du diviseur non biaisé $n - p$. Cette équation sert uniquement à la sélection des variables; une fois les variables choisies, on ajuste de nouveau le modèle à l'aide des MCO; voir équations (1) à (6) ci-dessus et FN pour une analyse.

Le raisonnement statistique n'est pas évident et il faut prendre les critères d'EKT au pied de la lettre. Par exemple, voici une autre équation possible:

$$\text{PEP 2-9} = .120 \text{ min} + .026 \text{ crim} + .029 \text{ class} - .176 \text{ pauv} + \text{résidu} \quad (12)$$

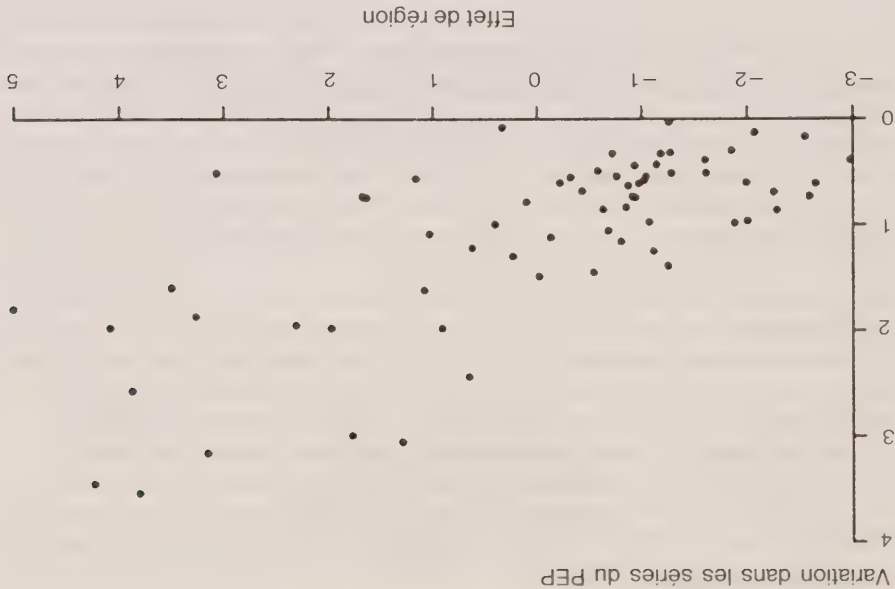
é.q.m. = 1.49.

(7.6) (3.4) (3.8) (-4.4)

La variable ajoutée est la proportion de personnes dans la région étudiée qui vivent au-dessous du seuil de pauvreté; le terme d'ordonnée à l'origine a été supprimé parce que la valeur *t* correspondante était peu élevée. L'équation (12) a un é.q.m. un peu meilleur que l'équation (11) et "montre" que le niveau de sous-dénombrement diminue lorsque la proportion de gens pauvres augmente, toutes choses étant égales par ailleurs. EKT rejettent cette équation sous prétexte que le coefficient de pauv est significativement négatif plutôt que significativement positif.

Les idées préconçues sur le sous-dénombrement peuvent être incompatibles avec les données, et la régression des MCO relative au meilleur sous-ensemble n'est peut-être pas une méthode analytique appropriée. Nous ne rejetons aucune des interprétations mais nous tirons la conclusion suivante: dans le contexte actuel, il n'existe pas de critère objectif et statistiquement défendable pour le choix des modèles. Ce choix repose en grande partie sur le jugement du modélisateur.

Figure 1. Le PEP et la qualité des données. Pour chacune des 66 régions étudiées, la graphique donne, sur l'axe horizontal, l'«effet de région» calculé par EKT et sur l'axe vertical, la variation dans les séries d'avril du PEP jugées supérieures.



L'association positive observée dans la figure 1 est tout à fait étonnante, de même que le changement que subit la distribution conjointe lorsque l'effet de région devient positif. Voici comment nous nous expliquons ces deux phénomènes: les estimations du sous-dénombrement tirées du PEP sont l'indice d'une faible qualité des données, dans le PEP comme dans le recensement. Un taux de sous-dénombrement manifestement élevé est l'indice d'une région où les données sont de piètre qualité. Les données manquantes sont nombreuses dans de telles régions; aussi, la modification des règles d'imputation aura des conséquences importantes. Les régions où il est difficile de dénombrer la population sont aussi des régions pour lesquelles il est difficile de redresser les chiffres du recensement. Voir FN, p. 9, ou Wolter (1986, p. 26, remarques 8 et 9).

On devrait pouvoir trouver une série du PEP – ou une combinaison formée à partir de ces séries – qui convienne. Mais pourquoi l'une ou l'autre de ces séries, ou une combinaison de celles-ci, seraient-elles une amélioration par rapport au recensement? Voilà la question cruciale, et EKT n'y apportent aucune réponse. À notre avis, tout redressement – qu'il se fasse au moyen d'une méthode synthétique, d'une série du PEP, d'un modèle de régression ou de n'importe quelle combinaison convexe – finit par reposer essentiellement sur des hypothèses.

7.2 Quelle série du PEP est la meilleure, et quelles variables explicatives utiliser?

À l'audience et dans leur analyse de FN, Erickson et Kadane préconisaient un redressement fondé sur la série PEP 2-9, celle qu'ils jugeaient supérieure à toutes les autres selon toute apparence. Nous avons choisi la série PEP 10-8 comme point de comparaison. EKT justifient l'utilisation de la série 2-9 et tentent d'éliminer quatre des douze séries, en particulier celle qui nous sert

leur dire, ils nous renvoient au tableau 11, qui indique une différence de taux de sous-dénombrement de 3.27 points pour la ville de New York, avec un degré d'incertitude de 0.62%. Dans beaucoup de circonstances, que des taux de sous-dénombrement soient uniformes importe peu; ce sont les différences de taux qui créent des inégalités. L'«effet de région» semble être un indice des différences de sous-dénombrement, variable qui suscite le plus grand intérêt. EKT ont calculé les «effets de région» indiqués dans le tableau de la façon suivante:

- i) ils ont concentré leur attention sur 8 des 12 séries du PEP;
- ii) ils ont lissé chacune des huit séries à l'aide du modèle de régression;
- iii) pour chaque région, ils ont fait la moyenne des huit estimations;
- iv) ils ont soustrait l'estimation du taux de sous-dénombrement national correspondante.

Dans le tableau 6 ci-dessous, nous comparons l'«effet de région» avec l'écart entre les estimations du PEP; nous limitons notre attention aux séries fondées sur la CPS d'avril et qui sont jugées supérieures. Les écarts qui peuvent exister entre les estimations du PEP sont uniquement attribuables à des différences dans le traitement des données manquantes. La série de chiffres de la troisième colonne semble raisonnable: les raisons qui font que des données sont manquantes peuvent varier selon les régions et il en va de même de la méthode d'imputation à utiliser. Si l'on ajoutait les séries d'août, l'intervalle de variation s'accroîtrait mais une partie de la différence serait attribuable à l'erreur d'échantillonnage.

Le tableau montre que, dans certains cas, l'effet de région est grand par rapport à l'écart entre les séries du PEP, ce qui laisse supposer que les données manquantes ont peu d'effet sur les résultats. C'est le cas du nord de l'Etat de New York. Pour d'autres régions toutefois, comme Chicago, on observe la relation inverse et les méthodes d'imputation ont alors de l'importance. Chacune des 66 régions étudiées est représentée par un point dans la figure 1. L'axe des x correspond à l'effet de région et l'axe des y, à la variation dans les séries d'avril du PEP jugées supérieures. Du point de vue de la moyenne quadratique (pour les 66 régions), l'écart entre les séries du PEP jugées supérieures par EKT, et fondées sur la CPS d'avril, équivalait à environ 75% de l'effet de région. Autrement dit, l'effet des données manquantes (abstraction faite des autres sources de biais présentes dans le PEP) est aussi grand que l'effet que tentent de mesurer EKT. L'introduction de nouveaux modèles d'imputation ne ferait qu'empirer les choses. La mise en moyenne ne serait pas non plus une bonne solution, pour les raisons évoquées plus haut.

Tableau 6
Comparaison de l'effet de région avec l'écart entre les estimations du PEP – séries jugées supérieures parmi celles fondées sur la CPS d'avril.
Les sous-régions correspondent à celles utilisées dans FN.

Effet de région	Séries d'avril du PEP jugées supérieures			région
	Min.	Max.	Variation	
Alabama	-.37	.60	.97	-1.07
Alaska	2.79	3.53	.74	1.63
Los Angeles	4.56	7.72	3.16	3.16
San Diego	-.98	1.45	2.43	.65
San Francisco	4.31	6.25	1.94	2.31
Reste de la Californie	2.84	3.92	1.08	1.03
Chicago	3.57	6.56	2.99	1.77
Reste de l'Illinois	1.21	1.75	.54	-1.04
New York City	6.04	7.90	1.86	3.27
Reste de l'Etat de New York	-1.61	-1.44	.17	-2.55
Wyoming	3.91	4.04	.13	1.16

EKT soutiennent que le PEP donne des estimations prudentes (p. 931). Cette affirmation semble à la fois erronée et hors de propos; erronée parce que les biais ont généralement pour effet de hausser artificiellement le sous-dénombrement, et hors de propos parce que la variation des biais dans l'espace a une grande importance. L'hypothèse (3) est peu vraisemblable: les erreurs n'ont probablement pas une moyenne nulle. Les taux de sous-dénombrement établis à l'aide des estimations du PEP risquent d'être entachés d'un biais par excès, la grandeur du biais variant selon la région. Pour une analyse des données pertinentes, voir Fay et coll., chap. 6; voir aussi FN. Dans la cause de 1980, le tribunal de première instance concluait:

«La preuve présentée au procès a permis d'établir que les estimations du PEP comportaient diverses erreurs attribuables à des lacunes dans la méthodologie du PEP. On désigne ces erreurs comme des "biais". La méthodologie du PEP est une importante source de biais car l'appariement des enregistrements de la CPS avec ceux du recensement [...] est une tâche extrêmement difficile et remplit d'incertitudes. Comme l'une et l'autre séries d'enregistrements pouvaient renfermer des données irrégulières, inexacts et incomplètes, le Bureau a commis de toute évidence de nombreuses erreurs en déterminant le code d'appariement des personnes dénombrées dans la CPS, ce qui a eu pour conséquence de fausser l'estimation du sous-dénombrement établie au moyen de l'échantillon P. En outre, la preuve présentée au procès a permis d'établir que les erreurs d'appariement étaient pour la plupart attribuables au fait que le Bureau avait tenu de nombreuses personnes pour oubliées dans le recensement alors que ce n'était pas le cas. Par conséquent, le PEP a donné lieu à une surestimation du sous-dénombrement. L'étendue de ce phénomène et sa variabilité d'une région à l'autre sont inconnues.» (TRADUCTION) (674 F Supp 1100, citations et renvois omis).

Examinons maintenant les équations (4) et (5). Tout d'abord l'hypothèse d'indépendance. En 1980, il y avait 3 centres d'exploitation et 12 bureaux régionaux. EKT répliquent qu'il y avait 400 bureaux de district. Soit. Mais il y avait aussi plusieurs douzaines de chefs de secteur, plusieurs centaines de milliers de préposés du recensement et quelque 1,500 intervieweurs pour la CPS. Les sources d'erreur sont nombreuses et il est fort probable que les erreurs ne soient pas indépendantes. Centres d'exploitation, bureaux régionaux, chefs de secteur, intervieweurs pour le recensement, intervieweurs pour la CPS: ce sont toutes là des composantes d'erreur, sans parler des répondants. De même, l'invariabilité de σ^2 dans (4) est peu vraisemblable: le sous-dénombrement n'a pas les mêmes causes dans toutes les régions du pays et une équation de régression linéaire ne suffit pas pour refléter toutes ces causes. Nous avons souligné que des événements fortuits comme une tempête de neige pouvaient engendrer une corrélation des erreurs dans plusieurs régions; EKT répliquent qu'il n'y a pas eu de tempête de neige. Cette question touche les fondements de la statistique: si les conditions atmosphériques sont bonnes, les erreurs sont indépendantes, mais si le temps est mauvais, tout est possible. Les distributions du modèle, de même que les inférences statistiques, dépendent donc de certains événements. Lesquels, et pourquoi? Heureusement, nous n'avons pas à nous pencher sur la différence entre l'inférence conditionnelle et l'inférence inconditionnelle. Un événement majeur est venu perturber les opérations du recensement dans plusieurs Etats de la région Pacific Northwest. En mai 1980, le mont St. Helens entra en éruption en plein durant la période des interviews de suivi.

7. AUTRES QUESTIONS

7.1 Importe-t-il d'utiliser une série plutôt qu'une autre?

Compte tenu du niveau de précision qu'EKT espèrent obtenir pour le recensement, les diverses séries du PEP – même celles qu'ils jugent supérieures – produisent des résultats vraiment différents entre eux, comme le montrent les équations (7), (8) et (9). Pourtant, EKT soutiennent que les séries qu'ils jugent supérieures donnent toutes des résultats semblables. Et pour appuyer

EKT poursuivent en décrivant divers plans pour l'échantillonnage par saisie-résaisie, laissant ainsi de côté la question du redressement pour petites régions pour 1990. En 1980, le litige portait en bonne partie sur la faisabilité d'un redressement pour les sous-régions des 66 régions étudiées. Pour gagner sa cause, New York devait démontrer qu'un tel redressement améliorerait les chiffres du recensement. Aujourd'hui, EKT semblent reconnaître qu'à cet égard, les éléments de preuve étaient insuffisants.

5. MISE EN MOYENNE ET ANALYSE DE SENSIBILITE

Les 12 séries du PEP sont le résultat d'une analyse de sensibilité portant sur les données manquantes. Comme la quantité de données manquantes était élevée par rapport au niveau de sous-dénombrement, les méthodes de traitement des données manquantes ont de l'importance. EKT proposent donc un éventail de méthodes pour redresser les chiffres du recensement en fonction des diverses séries du PEP; ces méthodes sont les suivantes: a) suppression des séries contradictoires (pp. 937-939); b) suppression des différences systématiques entre les séries (pp. 937-938); c) régression sur d'autres variables (l'estimateur "composite", p. 933 et suiv.); d) mise en moyenne (pp. 931 et 937).

Cette énumération met en évidence l'incertitude fondamentale des méthodes de redressement des chiffres du recensement et dans ce cas, il est nécessaire d'interroger l'utilisation de moyennes comme manière de réduire cette incertitude. Des choix arbitraires en matière de modélisation peuvent être défendables s'ils ont peu de conséquences - c'est le critère habituel de la robustesse. L'analyse de sensibilité (qui consiste à modifier les hypothèses et à observer les effets de cette modification sur les résultats) peut rejeter le critère de la robustesse. Toutefois, il est inutile de mettre en moyenne les résultats d'une analyse de sensibilité. Les diverses séries du PEP ne sont pas des mesures répétées du sous-dénombrement. L'important dans une série du PEP n'est pas la moyenne mais l'écart, car c'est l'écart (par exemple, celui entre les séries d'avril) qui illustre l'effet d'hypothèses de modélisation différentes sur une même série de données.

6. HYPOTHESES

EKT (p. 937) affirment que le modèle produit de meilleurs résultats que les estimations du PEP et que la méthode synthétique. Le modèle donne effectivement de meilleurs résultats que les estimations du PEP si on reconnaît la validité de ses hypothèses - équations (1) à (6) ci-dessus. Or, ces équations semblent encore très peu plausibles. De la même manière, le modèle donne de meilleurs résultats que la méthode synthétique à la seule condition qu'il utilise les variables additionnelles d'une manière raisonnable, ce qui nous ramène aux hypothèses.

Par moments, EKT semblent affirmer que l'on peut déduire le modèle des données (p. 933 et suiv.). Évidemment, la construction d'un modèle de régression ne se résume pas au choix des variables qui figureront dans les membres de l'équation, bien que cette opération soit assez difficile comme nous le verrons plus loin. De nombreuses questions surgissent à l'esprit: pourquoi les effets sont-ils linéaires et additifs? (équations (1) et (2) ci-dessus); que penser des hypothèses sur les erreurs? (équations (3), (4), (5) et (6)), et ainsi de suite. EKT ne présentent aucune preuve visant à appuyer leurs hypothèses, si ce n'est de tenter de répondre à nos réutations (p. 931). Croient-ils qu'un modèle est bon jusqu'à ce qu'on puisse démontrer le contraire?

De toutes manières, nous nous en tenons à nos propos. Pour des données sur le biais de corrélation, voir Fay et coll. (1988, section 6F en particulier); pour une critique des estimations d'Ericksen et Kadane, voir Fellegi (1985, p. 118). D'autres sources de biais dans les séries du PEP consistent dans les erreurs d'appariement et les erreurs de déclaration d'adresse.

de savoir si ce sont les chiffres bruts ou les chiffres redressés qui se rapprochent le plus de l'effectif de la population réelle hypothétique. En réalité, Schirm et Preston considèrent plusieurs distributions conjointes qui correspondent à divers "scénarios", c'est-à-dire à divers choix de paramètres; les résultats sont assez semblables d'un scénario à l'autre. Schirm et Preston considèrent aussi plusieurs fonctions de perte, ou mesures d'exacritude.

Nous commentons brièvement le scénario I.

- a) Le gain rapporté est plutôt modeste. Par exemple, les Etats auxquels ont profité les opérations de redressement regroupent en moyenne un peu plus de la moitié de la population totale, peu importe la petitesse du gain.
- b) La population "réelle" a été construite suivant l'hypothèse de la méthode synthétique, à savoir aucune variation systématique du taux de sous-dénombrement d'un groupe racial entre les régions; la variation aléatoire était permise. Voir l'équation (2) dans Schirm et Preston. Ainsi, la définition de "réalité" favorise le redressement synthétique.

Dans l'ensemble toutefois, l'argument de Schirm et Preston est raisonnable. Si les hypothèses de la méthode synthétique se vérifient plus ou moins, les estimations obtenues par cette méthode seront bonnes. Mais voilà: ces hypothèses se vérifient-elles? et quel genre de variation spatiale les taux de sous-dénombrement subissent-ils? Schirm et Preston n'apportent aucune réponse à ces questions. Dans la cause de 1980, le tribunal de première instance observe que "[...] la méthode synthétique fait tout simplement abstraction de la variation spatiale et suppose qu'un habitant de l'Alabama a autant de chances qu'un habitant de l'Alaska de ne pas être recensé. Toutefois, comme l'ont si bien exposé des témoins experts pour le défendeur, l'hypothèse selon laquelle les taux de sous-dénombrement pour les divers groupes d'âge-sexe et les divers groupes raciaux ne varient pas d'une région à l'autre n'a absolument aucun fondement [...] la méthode synthétique est clairement inadéquate pour redresser les chiffres du recensement." (TRANSCRIPTION) (674 F Supp 1098, citations et renvois omis).

4. REDRESSEMENT POUR LES PETITES RÉGIONS

Le redressement des chiffres du recensement est plus susceptible d'être profitable à des niveaux d'aggrégation géographique relativement élevés (par ex.: régions de recensement ou divisions de recensement). Or, il y a aux Etats-Unis 39,000 administrations publiques de divers niveaux (Etat, municipalité, etc.) qui réclament toutes des fonds publics. Bon nombre de ces administrations sont elles-mêmes décomposées en unités de niveau inférieur (par ex.: arrondissements, sections électorales, etc.). Si l'on doit redresser les chiffres du recensement, il faut le faire à un niveau de décomposition géographique assez élevé pour satisfaire à des exigences légales et administratives. De fait, pour le recensement de 1990, on a proposé de redresser les chiffres jusqu'au niveau de l'Etat. (L'Etat est la plus petite unité de la géographie du recensement; on en compte 6,5 millions aux E.-U.)

EKT étudie deux méthodes synthétiques ainsi qu'une méthode de régression pour redresser les chiffres du recensement dans des sous-régions des 66 régions étudiées (p. 941). Au bout du compte toutefois, rien ne prouve que le redressement pour petites régions a pour effet d'améliorer les chiffres bruts du recensement. Par rapport à 1980, EKT affirme (p. 943):

"Pour ce qui a trait aux 66 régions visées par notre étude, nous avons bon espoir d'améliorer les chiffres bruts du recensement, particulièrement dans les régions qui affichent un taux de sous-dénombrement ou de surdénombrement élevé et qui nécessitent le plus une opération de redressement. Nos résultats ne nous permettent pas de tirer des conclusions définitives quant aux régions suburbaines, aux villes centrales qui ne comptent pas parmi les 16 ayant fait l'objet de l'étude, ou aux autres régions urbaines ou rurales des Etats étudiés. Si on veut calculer des estimations pour ces régions, il serait préférable de ne pas reprendre les équations de régression présentées dans cet article." (TRANSCRIPTION)

Tableau 4
Taux de sous-dénombrement calculés d'après les chiffres du tableau 3, en pourcentage (les taux négatifs représentent en réalité un surdénombrement.)

Blancs		Noirs		Total
Etat A	- 1.1%	50%	17%	0%
	- 2.2%			0%
Etat B			17%	
Total	- 2.1%	17%		0%

Le contre-exemple a été élaboré de manière à permettre un calcul simple; nous aurions pu évidemment fournir un exemple plus complexe et plus réaliste. Selon le tableau 3, l'erreur globale dans le recensement (Blancs et Noirs) est nulle, pour chaque Etat comme pour le pays. Par conséquent, les chiffres du recensement reflètent les parts de population réelles de chaque Etat et toute tentative de redressement aurait des conséquences fâcheuses. Le tableau 4 donne les taux d'erreur (calculés en fonction de la population réelle); les conditions de Schirm et Preston sont respectées. Comme on peut le voir dans le tableau 5, le redressement synthétique fausse l'effectif total pour chaque Etat: un surdénombrement dans le cas de B et un sous-dénombrement dans le cas de A. Pour déduire le tableau 5 du tableau 3, on multiplie, par exemple, le nombre de Blancs dans l'Etat A par:

$$\text{nombre réel de Blancs au pays/nombre recensé} = 979/1,000. \tag{10}$$

On effectue le même calcul pour les autres cases.

Le contre-exemple peut nous aider à formuler le problème en termes de comparaison: l'Etat A a une faible densité de population et compte peu de membres de minorités; l'Etat B est fortement peuplé et compte une large minorité difficile à dénombrer. Le redressement synthétique peut favoriser des Etats comme B au détriment d'Etats comme A. L'erreur mathématique relevée dans l'annexe de Schirm et Preston semble être dans le raisonnement qui découle du schéma A.2. M. Preston nous dit (par communication personnelle) que le théorème se vérifie moyennant un ensemble de conditions plus complexe qui implique l'utilisation de moyennes pondérées.

Voilà pour notre critique de l'argument analytique de Schirm et Preston. Que penser maintenant de l'argument fondé sur la simulation? Essentiellement, Schirm et Preston considèrent 51 régions (les Etats plus le District de Columbia) et deux groupes raciaux (les Noirs et les Blancs). Ils définissent une distribution conjointe pour une population "réelle" hypothétique. Et pour les chiffres du recensement; dans les deux cas, il s'agit d'une distribution stochastique. Les chiffres du recensement peuvent être redressés à l'aide de la méthode synthétique et il s'agit

Tableau 5

Le redressement synthétique (Syn).

Blancs		Noirs		Total
Syn	Nombre réel	Syn	Nombre réel	Syn
Etat A	88	1	2	89
Etat B	891	120	119	1,011
Total	979	121	121	1,100

Les estimations du PEP s'accordent mieux avec les estimations synthétiques A dans le tableau 1. Mais cela est une relation tautologique car le taux de sous-dénombrement des per-
sonnes d'origine hispanique dans le modèle synthétique A a été estimé à l'aide des données
du PEP, tandis que le modèle synthétique B repose sur l'analyse démographique. Les diffé-
rences observées entre les estimations du PEP sont une réalité gênante; de même les différences
entre les estimations du PEP et les estimations synthétiques.

Voici la principale affirmation d'EKT (p. 927):

«Nous en venons à conclure que peu importe qu'on utilise l'une des méthodes simples
ou la méthode composite et peu importe la manière dont on modifie les hypothèses de
la méthode composite, un redressement réduit avec exactitude les parts de population
dans les Etats qui comptent une faible proportion de minorités et accroît les parts de
population dans les grandes villes.» (TRANSDUCTION)

Donner plus d'argent aux villes en corrigeant les chiffres du recensement est une idée louable,
mais seulement dans la mesure où on est sûr que les opérations de redressement amélioreront
la qualité des chiffres du recensement. La qualité des chiffres est la question au coeur du débat
et nous aurions aimé qu'EKT l'aborde plus directement. Leur tableau 5 a peu de rapport
avec la question.

3. SCHIRM ET PRESTON

Peut-on améliorer réellement la qualité des chiffres du recensement à l'aide d'un modèle
de redressement synthétique? EKT pensent que oui, citant Schirm et Preston (1987) à l'appui.
Schirm et Preston présentent deux arguments majeurs: l'un analytique et l'autre fondé sur la
simulation. Cependant, les deux ont des lacunes importantes.

L'argument analytique (p. 966):

«Notre conclusion est que le redressement synthétique produira toujours un rapport
estimé de la population d'un Etat à la population du pays plus près de la réalité si
a) le taux de sous-dénombrement des Noirs pour cet Etat se rapproche plus du taux de
sous-dénombrement des Noirs au niveau national que du taux de sous-dénombrement
en général;
b) le taux de sous-dénombrement des Blancs pour cet Etat se rapproche plus du taux de
sous-dénombrement des Blancs au niveau national que du taux de sous-dénombrement
en général.» (TRANSDUCTION)

Du point de vue mathématique, cette proposition est fausse. Nous fournissons un contre-
exemple dans le tableau 3: l'Etat A, par exemple, a une population blanche de 89 personnes
mais 90 ont été recensées.

Tableau 3

Contre-exemple infirmant l'argument analytique, avec deux Etats et deux groupes raciaux.

Blancs		Noirs		Total	
Nombre recensé	Nombre réel	Nombre recensé	Nombre réel	Nombre recensé	Nombre réel
Etat A	90	89	1	91	91
Etat B	910	890	99	1,009	1,009
Total	1,000	979	100	1,100	1,100

En abrégé,

é.q.m. entre les chiffres du recensement et les estimations synthétiques $B = 0.13$ de 1%.

EKT disent avoir une préférence pour les huit premières séries du PEP (pp. 933 et 938). Nous allons calculer maintenant l'é.q.m. entre la série 2-20 et la série 3-8, qui comptent parmi les séries jugées supérieures par EKT. (Les séries 2-20 et 3-8 reposent toutes deux sur la CPS d'avril; les différences qui peuvent exister entre elles sont uniquement attribuables à la méthode de traitement des données manquantes.)

é.q.m. entre les séries 2-20 et 3-8 du PEP = 0.14 de 1%.

EKT proposent aussi la mise en moyenne comme manière d'éliminer les imprécisions (pp. 931 et 937). Le tableau 2 compare les proportions de population établies d'après les chiffres du recensement avec celles calculées à l'aide des estimations synthétiques B ou de la moyenne des estimations du PEP jugées supérieures par EKT. Nous calculons l'é.q.m. entre la moyenne des estimations du PEP jugées supérieures et les estimations synthétiques B:

é.q.m. entre la moyenne des estimations du PEP jugées supérieures et les estimations synthétiques $B = 0.25$ de 1%.

La comparaison de (7), (8) et (9) nous amène à faire les observations suivantes:

- a) la différence entre les chiffres du recensement et les estimations synthétiques B est plutôt faible;
- b) l'intervalle de variation dans les séries du PEP jugées supérieures est plus grand que la différence entre les chiffres du recensement et les estimations synthétiques B;
- c) la différence entre la moyenne des estimations du PEP jugées supérieures et les estimations synthétiques B est le double de celle entre les chiffres du recensement et les estimations synthétiques B.

Un écart de 0.13% doit être important aux yeux d'EKT: voir (7). Si c'est le cas, les séries du PEP ne concordent pas entre elles. En outre, ces séries sont très différentes des estimations synthétiques. On pourrait certes prétendre, à ce sujet, que Schirm et Preston ne sont pas allés assez loin. Or, un comité de révision de la National Academy of Sciences – dont Jay Kadane est un membre important – en est venu à la conclusion provisoire que Schirm et Preston avaient déjà "surdressé" les chiffres du recensement; voir Cohen et Citro (1985, p. 287).

Tableau 2
Proportions de population établies d'après les chiffres du recensement, les estimations synthétiques B et la moyenne des huit séries d'estimations du PEP jugées supérieures par EKT (2-20, 3-20, 2-9, 3-9, 2-8, 3-8, 5-9, 5-8).

	Groupe 1	Groupe 2	Groupe 3	Total
Moyenne des séries du PEP jugées supérieures B	11.18%	44.34%	44.48%	100.00%
Synthétiques B	10.88%	44.30%	44.82%	100.00%
Recensement	10.76%	44.24%	45.00%	100.00%
Moyenne des séries du PEP jugées supérieures – Synthétiques B	-.30%	-.40%	-.34%	.00%
Recensement – Synthétiques B	-.12%	-.06%	+.18%	.00%

Tableau 1

Le tableau 5 d'EKT. Variation de parts de la population nationale lorsque les chiffres du recensement sont redressés à l'aide d'estimations d'échantillon pour diverses régions et d'estimations synthétiques. [Les chiffres des trois premières colonnes représentent des variations de parts, ou des différences de taux de sous-dénombrement; les chiffres de la dernière colonne représentent des taux de sous-dénombrement global.]

Estimations du PEP	Groupe 1	Groupe 2	Groupe 3	Taux de sous-dénombrement national estimé
--------------------	----------	----------	----------	---

2-20	+ .52%	+ .09%	-.61%	+ 1.9%
3-20	+ .51%	+ .08%	-.59%	+ 1.7%
2-9	+ .50%	+ .06%	-.56%	+ 1.6%
3-9	+ .49%	+ .04%	-.53%	+ 1.4%
2-8	+ .41%	+ .04%	-.45%	+ 1.1%
3-8	+ .39%	+ .03%	-.42%	+ 1.0%
5-9	+ .31%	+ .25%	-.56%	+ 2.1%
5-8	+ .22%	+ .23%	-.45%	+ 1.7%
14-20	+ .21%	+ .02%	-.23%	-.2%
10-8	+ .19%	+ .07%	-.26%	+ .3%
14-9	+ .19%	-.01%	-.18%	-.5%
14-8	+ .10%	-.03%	-.07%	- 1.0%
Synthétiques A	+ .17%	+ .14%	-.31%	+ 1.4%
Synthétiques B	+ .12%	+ .06%	-.18%	+ 1.4%
Proportion de l'effectif recensé	10.76%	44.24%	45.00%	

Notes: (i) Le groupe 1 comprend 16 villes centrales. Le groupe 2 comprend trois Etats amputés des villes du groupe 1 (Californie, Maryland et Texas) et 17 Etats entiers. Toutes les régions comptent au moins 10% de Noirs ou de personnes d'origine hispanique. Le groupe 3 comprend neuf Etats amputés d'une partie de leur territoire plus 21 Etats entiers. Toutes les régions du groupe 3 comptent moins de 10% de Noirs ou de personnes d'origine hispanique.

(ii) La méthode d'estimation synthétique A suppose que a) le taux de sous-dénombrement est le même pour les Noirs et pour les personnes d'origine hispanique, soit 5.9%; b) le taux de sous-dénombrement des personnes qui ne sont ni Noirs ni d'origine hispanique est de 0.3%; c) le taux de sous-dénombrement pour les Noirs, les personnes d'origine hispanique et tous les autres groupes ne varie pas selon les régions géographiques et d) il y a 3 millions d'étrangers sans document au pays, dont 9.6% sont des Noirs.

(iii) D'après Schirm et Preston (1987), la méthode d'estimation synthétique B suppose que a) le taux de sous-dénombrement chez les Noirs est de 5.9%; b) le taux de sous-dénombrement pour les personnes d'origine hispanique et les autres personnes non noires est de 0.7%; c) le taux de sous-dénombrement pour les Noirs, les personnes d'origine hispanique et tous les autres groupes ne varie pas selon les régions géographiques et d) il y a 3 millions d'étrangers sans document au pays, dont 9.6% sont des Noirs.

De plus, nous croyons que l'impression de conformité qui se dégage du tableau est très illusoire. Il y a des différences notables entre les séries du PEP jugées supérieures par EKT, ou entre ces séries et les estimations synthétiques de Schirm et Preston. Evidemment, la gravité de la chose tient à l'échelle de mesure; c'est pourquoi nous nous attachons maintenant à choisir les unités. Les partisans du redressement recourent souvent à la notion de "fonction de perte" pour soutenir leur idée; la plus courante est la fonction quadratique de perte: voir Ericksen, Estrada, Tukey et Wolter (1991, p. 20). EKT considèrent que Schirm et Preston ont démontré qu'il était avantageux de redresser les chiffres du recensement; nous allons donc calculer l'écart quadratique moyen (é.q.m.) entre les chiffres du recensement et les estimations synthétiques B (ligne "Synthétiques B" du tableau 1), obtenues à l'aide de la méthode de redressement de Schirm et Preston. (La moyenne est pondérée par la proportion de l'effectif recensé.)

$$\sqrt{.11 \times (.12)^2 + .44 \times (.06)^2 + .45 \times (.18)^2} \approx 0.13 \text{ de } 1\%$$

Dans (4), K_i est la variance d'échantillon fractionnée pour y_i calculée par le Bureau; on fait abstraction du caractère aléatoire de K_i ; σ^2 ne dépend pas de i et est tenue pour fixe même si elle est estimée à l'aide des données. Le rôle des hypothèses de même que les cas de non-vérification de ces hypothèses sont analysés dans FN; voir aussi les documents de travail et la réponse à la réplique, de même que les sections 6 et 7 ci-dessous.

Dans l'affaire de 1980, le modèle d'Erickson-Kadane avait servi à lisser les estimations du PEP dans le but de réduire l'erreur d'échantillonnage. L'objet principal de FN était de faire la critique de ce modèle. Erickson, Kadane et Tukey (1989) – que nous désignerons ci-dessous par EKT – ont répondu à FN. Le présent article se veut le prolongement du débat. EKT cite un article de Schirm et Preston (1987) dans lequel les auteurs envisagent le redressement des chiffres de population des Etats et du District de Columbia à l'aide de la méthode synthétique. Par exemple, l'analyse démographique (avec une série d'hypothèses sur l'immigration illégale) a permis d'établir un taux de sous-dénombrement national estimé de 5.9% pour les Noirs et de 0.7% pour les Blancs en 1980. La méthode synthétique corrige les chiffres de chaque Etat de la façon suivante: accroître le nombre de Noirs de 5.9% et le nombre de Blancs de 0.7%. Bref, on suppose que les taux de sous-dénombrement dépendent de l'origine raciale et non de la région géographique ni d'aucun autre facteur.

Cela termine notre description du contexte. Pour connaître les derniers développements en ce qui concerne le recensement de 1990, se référer à Freedman (1991); une partie des commentaires d'introduction du présent article sont extraits, à quelques différences près, de l'article de Freedman. Pour connaître d'autres points de vue, voir Hogan et Wolter (1988), Schirm (1991), Wolter (1991), Causey (1991) ou Erickson, Estrada, Tukey et Wolter (1991). Dans le reste de cet article, nous nous attachons à répondre aux principaux arguments d'EKT et à montrer comment le tribunal de première instance en est venu à trancher certaines questions.

2. LES REDRESSEMENTS AMÉLIORENT-ILS LES CHIFFRES DE RECENSEMENT?

La question fondamentale est de savoir si des redressements ont pour effet d'améliorer les chiffres de recensement. EKT ne doute pas "[...] de pouvoir améliorer les chiffres bruts du recensement [...]"] (p. 943) (TRADUCTION); en effet, il existe

"[...] deux méthodes simples (synthétiques) de redressement par lesquelles on peut améliorer les chiffres du recensement [...] en ce qui a trait au modèle d'Erickson et Kadane, la question n'est pas de savoir si ce modèle prouve qu'un redressement est possible mais de savoir s'il représente une amélioration par rapport aux méthodes plus simples (p. 927-928) [...] L'étude de la méthode ne "prouvera" pas qu'un redressement a pour effet d'améliorer les chiffres du recensement. Cela a déjà été confirmé par Schirm et Preston et les résultats des tableaux 5 et 6." (p. 933) (TRADUCTION)

Ainsi, les tableaux 5 et 6 d'EKT constituent l'essentiel de la preuve visant à montrer que le redressement améliorera les chiffres du recensement. Soit dit en passant, le tableau 6 sur les enregistrements erronés est redondant puisque les estimations du PEP contenues dans le tableau 5 tiennent compte de l'effet des enregistrements erronés. Le tableau 5 est donc le plus important et nous le reproduisons ici pour des raisons de commodité. À notre avis, ce tableau n'indique pas vraiment s'il est possible d'améliorer les chiffres du recensement; pour comprendre pourquoi, il nous faut quelques chiffres. (L'article de Schirm et Preston est analysé dans la section suivante.)

Dans le tableau ci-dessus, le "Groupe 1" se compose de 16 villes centrales, le "Groupe 2" comprend d'autres régions qui comptent une proportion relativement forte de minorités et le "Groupe 3" se compose de régions qui comptent une faible proportion de minorités. Au mieux, ce tableau montre que plusieurs méthodes de redressement peuvent produire des résultats généralement comparables. Il n'indique pas si l'une ou l'autre de ces méthodes accroît réellement l'exactitude des chiffres du recensement et avec raison, car il n'existe aucun critère externe permettant de le vérifier.

Dans l'affaire du recensement de 1980, Gene Erickson, Jay Kadane et John Tukey ont agi comme témoins experts pour les demandeurs. La méthode qu'ils préconisaient pour redresser les chiffres du recensement à l'aide des données du PEP est décrite dans Erickson et Kadane (1985). Freedman, parmi d'autres statisticiens et démographes, a témoigné pour les défenseurs et Navidi a agi à titre de consultant. Une critique de la méthode de redressement proposée par Erickson et Kadane est résumée dans Freedman et Navidi (1986), désigné ci-dessous par FN. Nous allons aborder maintenant quelques-uns des aspects techniques de la question. Selon des experts du Bureau of the Census:

- a) il y avait des différences notables entre les 12 séries du PEP, ce qui prouvait que les données manquantes représentaient un problème sérieux.
- b) les estimations du PEP pouvaient être entachées d'un biais appréciable, à part les problèmes créés par les données manquantes.
- c) chaque série du PEP était exposée à une erreur d'échantillonnage exagérément élevée.

Erickson et Kadane ont répliqué qu'une des séries du PEP ("PEP 2-9") était supérieure aux autres et que l'on pouvait réduire sensiblement l'erreur d'échantillonnage à l'aide de modèles de régression. Ils ont proposé un modèle à deux équations. La première exprime l'idée que y_i , l'estimation du PEP pour la région i , est une estimation sans biais du sous-dénombrement réel, y_i , pour cette région. Ainsi,

$$\text{Estimation du PEP pour la région } i = \text{sous-dénombrement réel dans la région } i + \text{erreur aléatoire.}$$

En termes mathématiques,

$$(1) \quad y_i = \gamma_i + \delta_i$$

La seconde équation exprime une théorie sur la variation du sous-dénombrement entre les régions en fonction d'un vecteur de variables explicatives, X_i , et d'un vecteur d'hyperparamètres, β . En termes généraux,

$$\text{Sous-dénombrement réel dans la région } i = \text{combinaison linéaire de variables explicatives pour la région } i + \text{erreur aléatoire.}$$

En termes mathématiques,

$$(2) \quad y_i = X_i \cdot \beta + \epsilon_i$$

Les hypothèses relatives aux termes d'erreur peuvent être formulées comme suit:

$$(3) \quad E(\delta_i) = E(\epsilon_i) = 0.$$

$$(4) \quad \text{var} \delta_i = K_i, \text{ var} \epsilon_i = \sigma^2.$$

$$(5) \quad \delta_1, \delta_2, \dots, \delta_{66}, \epsilon_1, \epsilon_2, \dots, \epsilon_{66} \text{ sont indépendantes.}$$

$$(6) \quad \delta_i \text{ et } \epsilon_i \text{ sont distribuées selon une loi normale.}$$

soutient qu'un redressement introduira des inexactitudes encore plus grandes dans les chiffres de population et qu'en conséquence, une telle opération n'est pas techniquement réalisable ou justifiable à l'heure actuelle." (TRADUCTION) (674 *F Supp* 1091, c.-à-d. volume 674 du *Federal Supplement*, page 1091).

L'affaire du recensement de 1980 peut sembler avoir perdu de son intérêt étant donné que le recensement de 1990 a déjà eu lieu. Cependant, de toutes les causes où il a été question de principes statistiques, celle du recensement de 1980 compte parmi l'une des plus importantes et des plus ardemment défendues; il y a encore beaucoup de leçons à tirer de cette affaire. Dans cet article, nous revenons sur quelques aspects techniques de la question et sur certaines conclusions du tribunal.

Le reste de cette section sert à décrire le contexte du débat; pour plus de détails, se référer à Cohen et Citro (1985) ou Fay et coll. (1988). Deux méthodes permettent d'évaluer le sous-dénombrement dans les recensements aux E.-U.: l'analyse démographique et la saisie-résaisie. En analyse démographique, on se sert de dossiers administratifs (certificats de naissance, certificats de décès, visas d'immigrant, etc.) pour établir des estimations indépendantes de totaux de population. Le tout repose sur une identité fondamentale:

$$\text{Population} = \text{naissances} - \text{décès} + \text{immigration} - \text{émigration}.$$

L'analyse démographique produit des estimations selon l'âge, le sexe et l'origine raciale mais non selon l'origine ethnique, à cause du manque de renseignements dans les dossiers. Les données sur l'immigration et l'émigration sont incomplètes; les registres des naissances aussi, surtout en ce qui a trait à la période antérieure à 1935. Il faut donc compléter les données qui entrent dans l'identité fondamentale par une série d'imputations et de corrections. En outre, comme il existe peu de données sur la migration interne, les estimations pertinentes sont calculées surtout au niveau national. Voilà pour ce qui a trait à l'analyse démographique.

L'estimation du taux de couverture pour les petites régions (y compris les États et les villes) repose sur les techniques de saisie-résaisie. La saisie correspond au recensement; la résaisie correspond à une enquête par sondage effectuée après le recensement. En 1980, il y a eu deux enquêtes de ce genre, dites "échantillons *P*"; la CPS (Current Population Survey) d'avril et celle d'août. Chaque enregistrement de l'échantillon *P* est comparé aux fiches du recensement pour vérifier si la personne en question a été "saisie", c'est-à-dire dénombrée. Si un enregistrement ne peut être apparié avec aucune fiche, cela signifie que la personne en question a été "oubliée" ou encore, que le processus d'appariement ne fonctionne pas bien. Les données qui découlent de cet exercice servent à estimer le pourcentage de personnes qui ont été oubliées dans le recensement, c'est-à-dire le taux de sous-dénombrement.

Les fiches du recensement comprennent aussi un faible pourcentage d'enregistrements erronés (par exemple, des personnes qui ont été recensées à deux endroits différents); on estime le nombre de ces enregistrements en prélevant un échantillon d'enregistrements du recensement, appelé "échantillon *D*", que l'on soumet à une vérification sur le terrain. En définitive, on estime le sous-dénombrement net en déduisant le nombre d'enregistrements erronés du nombre de personnes oubliées. (Pour plus de détails, voir Fay et coll., chapitre 5.) L'estimation du sous-dénombrement s'inscrit dans le cadre du Post Enumeration Program (PEP). En 1980, le nombre de données manquantes était assez considérable dans les échantillons *P* et *D*. Par exemple, le taux de non-réponse dans la CPS était de 4%. En outre, on ne pouvait déterminer de code d'appariement pour 4% des personnes interviewées. Diverses techniques d'imputation ont été envisagées pour évaluer l'effet des données manquantes; on a obtenu ainsi 12 séries d'estimations du PEP pour 66 sous-régions.

Les 66 sous-régions étaient réparties sur tout le territoire des E.-U. Elles comprenaient des villes comme New York, des portions d'État, comme le nord de l'État de New York, et des États entiers, comme le Wyoming. Les "séries" du PEP comprennent chacune 66 estimations, une pour chaque région étudiée; neuf des douze séries étaient fondées sur la CPS d'avril et les trois autres, sur la CPS d'août.

Aurions-nous dû redresser les chiffres du recensement des E.-U. de 1980?

D.A. FREEDMAN et W.C. NAVIDI¹

RÉSUMÉ

Cet article examine quelques-uns des arguments invoqués à l'appui et à l'encontre du redressement des chiffres du recensement des E.-U. de 1980 et revoit la décision du tribunal.

MOTS CLÉS: Recensement; redressement; enquête postcensitaire; régression; lissage.

1. INTRODUCTION

À tous les dix ans, le recensement trace le portrait statistique de la population des États-Unis. Le niveau de décomposition géographique des données du recensement en fait des données uniques. Mais ces chiffres n'ont pas qu'un intérêt théorique: ils ont une influence sur la répartition du pouvoir et des ressources monétaires. Les résultats du recensement servent à la répartition des sièges au Congrès et dans les assemblées législatives locales et à la répartition des recettes fiscales – 40 milliards de dollars par année à la fin des années 1980 – entre 39,000 administrations publiques de divers niveaux (États, municipalités, etc.). Pour cette raison, la répartition géographique de la population a plus d'importance que les chiffres de population pour l'ensemble du pays. En effet, le recensement sert de base à la répartition d'une quantité fixe de ressources: si une administration publique reçoit plus, une autre recevra nécessairement moins. Le redressement des chiffres du recensement n'est recommandé que dans la mesure où cette opération aboutit à une représentation plus juste de la répartition de la population.

On relève toujours un certain niveau de sous-dénombrement, modeste, dans les recensements et ce sous-dénombrement est rarement uniforme. Les personnes qui déménagent au moment du recensement sont difficiles à dénombrer; dans les régions rurales, les cartes et les listes d'adresses sont incomplètes. Les villes centrales renferment de fortes concentrations de personnes défavorisées et de minorités, qui peuvent être plus difficiles à recenser. Si le sous-dénombrement peut être estimé avec assez de précision, surtout au niveau local, on peut envisager – et même recommander – des opérations de redressement en vue d'améliorer les chiffres du recensement. Certains statisticiens soutiennent qu'il est possible d'estimer assez précisément le sous-dénombrement; d'autres sont sceptiques sur la question: un mauvais redressement peut être pire que pas de redressement du tout.

En raison de ses effets sur la répartition des ressources, le sous-dénombrement a suscité beaucoup d'intérêt dans les médias, au Congrès et devant les tribunaux. À la suite du recensement de 1980, la ville de New York s'est jointe à d'autres administrations pour intenter une action contre le Département du commerce dans le but de l'obliger à effectuer un redressement basé sur l'analyse démographique et les techniques de saisie-resaisie. Le Département du commerce est resté sur ses positions. Le tribunal de première instance a exposé l'affaire dans les termes suivants:

«Les demandeurs soutiennent qu'un redressement des chiffres du recensement aura pour effet d'accroître le degré d'exacritude du recensement, rabaisant ainsi le sous-dénombrement excessif dans la ville et l'État [de New York]. De son côté, tout en reconnaissant que les chiffres du recensement ne sont pas parfaits, le Census Bureau

¹ D.A. Freedman, Département de statistique, Université de Californie, Berkeley, CA 94720; W.C. Navidi, Département de mathématiques, Université Southern California, Los Angeles, CA E.-U. 90089.

Roe, Carlson et Swanson décrivent une variante de la méthode des unités de logement pour estimer la population de petites régions rurales. Selon cette variante, des experts locaux fournissent des données sur des ménages échantillonnés. On compare les estimations pertinentes aux chiffres du recensement pour trois collectivités rurales.

Xia et coll. comparent les propriétés statistiques de l'échantillonnage par grappes à un seul degré étagé et de l'échantillonnage par grappes à un seul degré ordinaire de même que les coûts relatifs à chacun. On peut recourir au premier mode d'échantillonnage lorsqu'il n'est pas possible de sous-échantillonner des grappes (c.-à-d. que l'échantillonnage par grappes à deux degrés est impossible). La méthode a été appliquée dans l'enquête de Shanghai sur la maladie d'Alzheimer et la démence. On montre que l'on peut réduire les coûts sans sacrifier la précision.

Le rédacteur en chef

Dans ce numéro

Malgré le soin extrême que prennent les organismes statistiques à recenser une population, il subsiste toujours un certain niveau de sous-dénombrement. De plus, ce niveau varie habituellement d'un sous-groupe de la population à l'autre et ne se répercute donc pas également sur tous les programmes publics reposant sur les chiffres du recensement. C'est pourquoi les responsables de l'action gouvernementale et les statisticiens s'intéressent grandement aux méthodes de calcul du sous-dénombrement, aux techniques de redressement, particulièrement celles concernant les petites régions, et aux questions connexes. Les six articles qui constituent la section spéciale **Méthodes et questions concernant la mesure du sous-dénombrement du recensement** seront une précieuse contribution à la série d'ouvrages déjà nombreux sur le sujet.

Le premier article de la section spéciale est un document de travail de Freedman et Navidi. Les auteurs examinent quelques-uns des arguments statistiques invoqués à l'appui et à l'encontre du redressement des chiffres du recensement des États-Unis de 1980 et analysent la preuve statistique qui a été présentée au cours d'un procès intenté contre le Département du commerce et le Bureau of the Census des E.-U. Cet article est, en fait, un nouvel épisode du débat qui oppose les auteurs à Erickson, Kadane et Tukey, qui proposent des méthodes de redressement. Dans cet article, on montre aussi comment le tribunal de première instance a tranché quelques-unes des questions en litige. L'article est suivi de commentaires très pointus et vigoureux de plusieurs statisticiens ainsi que d'une réponse des auteurs.

Cressie présente une méthode empirique de Bayes pour la prédiction du sous-dénombrement à des niveaux intra-nationaux. Cette méthode est fondée sur l'estimation du maximum de vraisemblance avec contrainte (MVC). L'avantage des estimateurs MVC par rapport aux estimateurs du maximum de vraisemblance ordinaire est, dit-on, qu'ils ne tendent pas à lisser outre mesure les données des enquêtes postcensitaires. À l'aide d'un exemple et d'une simulation, Cressie compare l'estimateur MVC à l'estimateur du maximum de vraisemblance et à l'estimateur de la méthode des moments.

Avant le recensement de 1990 aux États-Unis, on a effectué une répétition générale dans l'État du Mississippi. Datta et coll. se servent des données de cet exercice pour étudier des méthodes de modélisation fondées sur des enquêtes postcensitaires. Ils s'intéressent aussi bien à des méthodes hiérarchiques de Bayes qu'à des méthodes empiriques de Bayes. Les résultats de leur analyse indiquent que les deux types de méthodes représentent une amélioration par rapport à l'estimation de système dual. Les auteurs terminent leur article par une mise au point rendue nécessaire par suite du redressement des chiffres du recensement de 1990.

Royce analyse quatre estimateurs de l'effet de la population de référence du Programme des estimations démographiques de Statistique Canada. Il s'agit de l'estimateur fondé sur les chiffres non redressés du recensement, de l'estimateur fondé sur les chiffres redressés du recensement, de l'estimateur d'essai préliminaire et de l'estimateur de comparaison pour ces estimateurs et s'applique moyennement pondérée (EQMP) sert d'élément de comparaison pour ces estimateurs et s'applique non seulement aux totaux de population, mais aux fonctions de totaux de population, comme les proportions de population, les taux de croissance, etc.

Swain et coll. décrivent dans ses grandes lignes le registre des adresses qui a été créé à Statistique Canada dans le but de réduire le niveau de sous-dénombrement dans le recensement de 1991 au Canada; ce registre constitue aussi une base de sondage des adresses domiciliaires pour les grands centres urbains et les agglomérations de taille moyenne. Les auteurs parlent de méthodologie, d'évaluation postcensitaire et de perspectives d'avenir.

Dans le dernier article de la section spéciale, Fienberg présente une bibliographie sélectionnée et commentée sur l'estimation de la taille de populations par la méthode de saisie-résaisie. L'estimation par saisie-résaisie est la principale méthode d'évaluation de l'intégrité du dénombrement; c'est pourquoi l'article insiste sur les ouvrages qui concernent l'estimation de populations humaines.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 18, numéro 1, juin 1992

TABLE DES MATIÈRES

Dans ce numéro	1
Méthodes et questions concernant la mesure du sous-dénombrement du recensement	
D.A. FREEDMAN et W.C. NAVIDI	
Aurions-nous dû redresser les chiffres du recensement des E.-U. de 1980?	3
Commentaires: S.E. FIENBERG	27
I.P. FELLEGI	31
N. CRESSIE	34
A.L. SCHIRM et S.H. PRESTON	37
J.A. HARTIGAN	47
T.P. SPEED	55
E.P. ERIKSEN et J.B. KADANE	57
Réponse des auteurs	65
N. CRESSIE	
Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des	
taux de sous-dénombrement du recensement selon l'approche empirique de Bayes .	83
G.S. DATTA, M. GHOSH, E.T. HUANG, C.T. ISAKI, L.K. SCHULTZ et J.H. TSAY	
Méthode hiérarchique de Bayes et méthode empirique de Bayes pour le redressement	
du sous-dénombrement: données de la "répétition générale" du recensement, effectuée	
en 1988 au Missouri	105
D. ROYCE	
Une comparaison d'estimateurs d'un ensemble de totaux de population	121
L. SWAIN, J.D. DREW, B. LAFRANCE et K. LANCE	
La création d'un registre des adresses résidentielles pour améliorer la couverture du	
recensement du Canada de 1991	139
S.E. FIENBERG	
Bibliographie sur la modélisation à l'aide de la saisie-resaisie avec application au	
redressement des chiffres du recensement pour éliminer le sous-dénombrement	157
L.K. ROE, J.F. CARLSON et D.A. SWANSON	
Une variante de la méthode des unités de logement pour estimer la population de petites	
régions rurales: une étude de cas portant sur la procédure des experts locaux	171
Z. XIA, P.S. LEVY, E.S.H. YU, Z. WANG et M. ZHANG	
Echantillonnage par grappes à un seul degré dans les enquêtes prévalence-incidence:	
certaines questions soulevées par l'enquête de Shanghai sur la maladie d'Alzheimer et	
la démence	181

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

- Président**
G.J. Brackstone
- Membres**
B.N. Chinnappa
G.J.C. Hole
F. Mayda (Directeur de la production)
R. Platek (Ancien président)
D. Patrick
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

- Rédacteur en chef**
M.P. Singh, *Statistique Canada*
- Rédacteurs associés**

- D.R. Bellhouse, *U. of Western Ontario*
D. Binder, *Statistique Canada*
E.B. Dagum, *Statistique Canada*
J.-C. Deville, *INSEE*
D. Drew, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*

- Rédacteurs adjoints**
P. Lavallée, L. Mantel, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

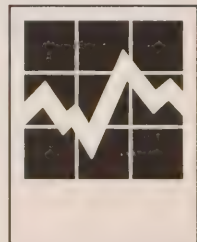
Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (É.-U.) aux États-Unis, et de 49 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

Statistique Canada
Division des méthodes d'enquêtes sociales

Techniques d'enquête

Une revue de Statistique Canada

Juin 1992 Volume 18 Numéro 1



Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1992

Tous droits réservés. Il est interdit de reproduire ou de
transmettre le contenu de la présente publication, sous quelque
forme ou par quelque moyen que ce soit, enregistré ou non,
support magnétique, reproduction électronique, mécanique,
photographique, ou autre, ou de l'emmagasiner dans un système
de recouvrement, sans l'autorisation écrite préalable du Chef,
Services aux auteurs, Division des publications, Statistique
Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 1992

Prix : Canada : 35 \$

États-Unis : 42 \$ US

Autres pays : 49 \$ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa



Catalogue 12-001

Techniques d'enquête

Une revue de Statistique Canada

Juin 1992 Volume 18 Numéro 1



Statistique
Canada

Statistics
Canada

Canada



Catalogue 12-001

Survey Methodology

A Journal of Statistics Canada

December 1992 Volume 18 Number 2



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

December 1992 Volume 18 Number 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1992

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 1992

Price: Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members B.N. Chinnappa C. Patrick
G.J.C. Hole D. Roy
F. Mayda (Production Manager) M.P. Singh
R. Platek (Past Chairman)

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>U. of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	L.-P. Rivest, <i>Laval University</i>
J.-C. Deville, <i>INSEE</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
R.E. Fay, <i>U.S. Bureau of the Census</i>	C.E. Särndal, <i>University of Montreal</i>
W.A. Fuller, <i>Iowa State University</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
J.F. Gentleman, <i>Statistics Canada</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	J. Sedransk, <i>State University of New York</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	C.M. Suchindran, <i>University of North Carolina</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat Inc.</i>
G. Kalton, <i>University of Michigan</i>	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

P. Lavallée, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 18, Number 2, December 1992

CONTENTS

In This Issue	177
Inference with Survey Data	
R.M. ROYALL Robustness and Optimal Design Under Prediction Models for Finite Populations..	179
T.M.F. SMITH and E. NJENGA Robust Model-Based Methods for Analytic Surveys	187
J.N.K. RAO, C.F.J. WU and K. YUE Some Recent Work on Resampling Methods for Complex Surveys	209
H.J. MANTEL An Estimating Function Approach to Finite Population Estimation	219
A.M. KRIEGER and D. PFEFFERMANN Maximum Likelihood Estimation from Complex Sample Surveys	225
C.-E. SÄRNDAL Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used.....	241
<hr/>	
J.B. ARMSTRONG and C.F.J. WU A Sample Allocation Method for Two-Phase Survey Designs	253
M.P. COUPER and R.M. GROVES The Role of the Interviewer in Survey Participation	263
P. LAHIRI and W. WANG A Multivariate Procedure Towards Composite Estimation of Consumer Expenditure for the U.S. Consumer Price Index Numbers	279
Acknowledgements	293

In This Issue

In August of 1991 a symposium in honour of Professor V.P. Godambe on the occasion of his 65th birthday was held at the University of Waterloo. Papers presented at this symposium were in the areas of foundations of inference, theory of estimation, and theory of survey sampling, all areas in which Professor Godambe has an interest and to which he has made significant contributions. The special section **Inference with Survey Data** in this issue, which is dedicated to Professor Godambe, contains some of the sampling related papers from the symposium. As a group these papers discuss many important issues for inference with survey data such as the role of modelling, robustness, complex survey designs, resampling methods, and the effects of imputation.

Royall considers model based estimation for finite population parameters. He describes the conflict between designs which provide model efficiency and those which are robust to model failure. Robustness is achieved through balanced samples. He presents a class of models for which the optimal sample is already balanced so that, for models in that class, there is no conflict between robustness and efficiency.

Smith and Njenga discuss model based and randomization based inference for sample surveys and suggest a robust non-parametric modelling approach to inference. Based on simulations using both real and synthetic data, they conclude that their estimator of a regression coefficient is robust to violations of assumptions of linearity and homoscedasticity, has good efficiency, and has reasonable conditional and unconditional properties.

Rao, Wu, and Yue review recent developments in resampling methods for complex survey designs, particularly the jackknife, balanced repeated replication, and the bootstrap. In a simulation study using a synthetic population they evaluate and compare variance estimators and confidence intervals for the population median.

Mantel considers model assisted estimation of a finite population mean based on a sample survey. He suggests that models should be extended so that the finite population mean is a known function of the optimal census based estimate of a model parameter. The extended model is then a compromise between model efficiency and finite population relevance.

Krieger and Pfeiffermann discuss maximum likelihood estimation of model parameters. They describe various approaches in the literature and consider the problem of informative designs. They propose the use of weighted distributions where the weights are modelled as functions of the covariates and of the variable of interest. The approach performs reasonably well in a simulation study.

In the final paper of this special section Särndal considers the problem of variance estimation when imputation is used to complete a data set. Overall variance is derived as the sum of a sampling variance and an imputation variance. The suggested variance estimator is a design based estimator of the sampling variance with a model based correction for bias and a model based estimator of the imputation variance. Some examples and an empirical evaluation are presented.

Armstrong and Wu formulate the problem of sample allocation for a general two-phase survey design as a constrained programming problem. By exploiting its mathematical structure, they propose a solution that consists of iterations between two subproblems that are computationally much simpler. They provide empirical results showing that the proposed method works very well.

Couper and Groves examine whether experienced interviewers achieve higher response rates than inexperienced interviewers, controlling for differences in survey design and attributes of the population assigned to them. After demonstrating that the relationship is positive and curvilinear, they attempt to explain the mechanisms by which experienced interviewers achieve these rates and elaborate the nature of the relationship.

Lahiri and Wang propose new estimators for the “cost weights” and “relative importances” which are needed to construct the U.S. Consumer Price Index Numbers. The proposed estimators are composite estimators that combine information from relevant sources. A numerical comparison with four rival estimators is also presented.

Robustness and Optimal Design Under Prediction Models for Finite Populations

RICHARD M. ROYALL¹

ABSTRACT

In many finite population sampling problems the design that is optimal in the sense of minimizing the variance of the best linear unbiased estimator under a particular working model is bad in the sense of robustness – it leaves the estimator extremely vulnerable to bias if the working model is incorrect. However there are some important models under which one design provides both efficiency and robustness. We present a theorem that identifies such models and their optimal designs.

KEY WORDS: Balanced sample; Bias protection; Model failure; Working model.

1. INTRODUCTION

The “ratio estimator” of a finite population total $T = y_1 + \dots + y_N$ is $\hat{T} = N\bar{x}\bar{y}_s/\bar{x}_s$, where $\bar{x} = (x_1 + \dots + x_N)/N$ is the known population mean of an auxiliary variable and \bar{x}_s and \bar{y}_s are sample means. This is the best linear unbiased (BLU) estimator of T under the model M :

$$E(Y_i) = \beta x_i,$$

$$\text{cov}(Y_i, Y_j) = \begin{cases} \sigma^2 x_i & i = j \\ 0 & \text{else.} \end{cases}$$

This estimator is biased under alternative models having different regression functions, in general, but protection against bias under specific alternatives can be assured by careful choice of the sample, as will be described below.

Throughout this paper we will be concerned with populations for which a particular model, such as M , is believed to apply, at least to a satisfactory degree of approximation. Our inferences will be made with reference to this model. For example, we will call an estimator \hat{T} unbiased only if $E_M(\hat{T} - T) = 0$. On the other hand, we recognize that the model is an approximation and that it might be seriously wrong. Thus we describe it as a **working model**, and seek sampling and estimation procedures that are robust in the sense of performing well, not only under that working model, but also under alternative models that might better describe the relationships between variables in our population.

We denote by $M(\delta_0, \delta_1, \dots, \delta_J : v)$ the general polynomial regression model:

$$E(Y_i) = \sum_{j=0}^J \delta_j \beta_j x_i^j$$

¹ Richard M. Royall, Johns Hopkins University, Baltimore, MD 21205 U.S.A.

$$\text{cov}(Y_i, Y_j) = \begin{cases} v_i \sigma^2 & i = j, \\ 0 & \text{else} \end{cases}$$

where δ_j is a zero-one indicator of whether the regressor x^j is included in the model. The best linear unbiased estimator under this model is denoted by $\hat{T}(\delta_0, \dots, \delta_J : v)$. Thus our first model was $M(0, 1 : x)$, and $\hat{T}(0, 1 : x)$ is the ratio estimator.

Royall and Herson (1973) showed that $\hat{T}(0, 1 : x)$ remains unbiased under $M(\delta_0, \dots, \delta_J : v)$ for any vector $(\delta_0, \dots, \delta_J)$ of zeroes and ones, and any v_1, \dots, v_N , if the sample is **balanced** on x, x^2, \dots, x^J :

$$\sum_s x_i^j / n = \sum_1^N x_i^j / n \quad j = 1, 2, \dots, J.$$

This means that in a balanced sample $\hat{T}(0, 1 : x)$ is robust in the sense that it remains unbiased under regression models that are much more general than the working model $M(0, 1 : x)$. Royall and Herson (1973, sec. 4.5) also detailed how approximate balance ensures the approximate unbiasedness of $\hat{T}(0, 1 : x)$. Furthermore they showed that in a balanced sample this estimator retains not only its unbiasedness but also its **optimality** under a wide variety of polynomial regression models, including $M(1 : 1)$, $M(1, 1 : x)$, and $M(0, 1, 1 : x^2)$. Specifically, the estimator is optimal under any polynomial regression model of degree J or less, provided only that the model's variance function is expressible as a linear combination of the regressors.

The robustness of the ratio estimator in balanced samples is achieved at a high cost in efficiency under the original working model $M(0, 1 : x)$. Under this model the sample that minimizes the variance consists of the n units whose x -values are largest, and the efficiency of a balanced sample is only $\bar{x} / \max_s(x_s)$. (Royall and Herson 1973).

For the linear regression estimator, theoretical results have been established that are quite analogous to those sketched above for the ratio estimator, but with one important difference. The estimator is $\hat{T}(1, 1 : 1) = N[\bar{y}_s + b(\bar{x} - \bar{x}_s)]$, where $b = \sum_s (x_i - \bar{x}_s) y_i / \sum_s (x_i - \bar{x}_s)^2$. It is the optimal (BLU) estimator under the constant variance linear regression model, $M(1, 1 : 1)$. When the sample is balanced, this estimator is robust, remaining unbiased (and optimal) under the same broad class of polynomial regression models as the ratio estimator. But unlike the ratio estimator, the regression estimator achieves robustness in balanced samples at **no cost in efficiency** – the variance under the working model $M(1, 1 : 1)$ is minimized in balanced samples, where $\bar{x}_s = \bar{x}$. This phenomenon occurs because the error variance $E(\hat{T} - T)^2$ is the sum of a constant and a term proportional to $(\bar{x} - \bar{x}_s)^2 \text{var}(b)$. Minimizing $\text{var}(b)$ requires maximizing $\sum_s (x_i - \bar{x}_s)^2$, but this term is eliminated altogether in samples with $\bar{x}_s = \bar{x}$.

Are there other models under which the same sample that minimizes the variance of the BLU estimator can also protect against bias under a wide range of alternative models? In particular, are there such models for problems requiring non-constant variance functions? We show that the answer is positive, giving a theorem that characterizes a family of models with the desired property and identifies the corresponding optimal samples. The results in this paper integrate and generalize those of Kott (1984) and Tallis (1986). They are also closely related to the work of Pereira and Rodrigues (1983) and Tam (1986), as well as that of Isaki and Fuller (1982).

2. BASIC RESULTS

It is convenient to shift to vector and matrix notation, in which Y is the population vector $(Y_1, Y_2, \dots, Y_N)'$ and the model $M(X: V)$ specifies that $E(Y) = X\beta$ and $\text{var}(Y) = V\sigma^2$, where X is an $N \times p$ matrix of regressors, V is diagonal, and the vector β and the scalar σ^2 are unknown. For a given sample s of n units we list the sample units first, so that

$$Y = \begin{pmatrix} Y_s \\ Y_r \end{pmatrix}, \quad X = \begin{pmatrix} X_s \\ X_r \end{pmatrix}, \quad V = \begin{pmatrix} V_s & 0 \\ 0 & V_r \end{pmatrix},$$

where Y_r is the $(N - n)$ -vector corresponding to the non-sample units, *etc.* We let 1_s and 1_r denote vectors $(1, \dots, 1)'$ of lengths n and $(N - n)$.

The population total is $T = 1_s'Y_s + 1_r'Y_r$. After the sample s is observed, the first component, $1_s'Y_s$, is known. The BLU estimator of T is obtained by adding to this known quantity the BLU predictor of $1_r'Y_r$:

$$\hat{T}(X: V) = 1_s'Y_s + 1_r'X_r\hat{\beta}(X: V),$$

where $\hat{\beta}(X: V) = (X_s'V_s^{-1}X_s)^{-1}X_s'V_s^{-1}Y_s$. The error variance is

$$\text{var}(\hat{T}(X: V) - T) = 1_r'(X_r'A_s^{-1}X_r + V_r)1_r\sigma^2,$$

where $A_s = X_s'V_s^{-1}X_s$. These formulas simplify when the vector $V1$ is in the linear manifold generated by the columns of X , which we denote by $\mathfrak{M}(X)$.

Lemma 1. If $V1 \in \mathfrak{M}(X)$ then

$$\hat{T}(X: V) = 1'X\hat{\beta}(X: V)$$

and under $M(X: V)$

$$\text{var}(\hat{T}(X: V) - T) = (1'XA_s^{-1}X'1 - 1'V1)\sigma^2.$$

Proof: The estimator simplifies because $V1 \in \mathfrak{M}(X)$ means that $V1 = Xc$ for some vector c , so that $X_s'1_s = X_s'V_s^{-1}X_sc$, from which we have $1_s'X_s\hat{\beta} = c'X_s'V_s^{-1}Y_s = 1_s'Y_s$. The variance formula follows from $\text{cov}(\hat{T}, T) = \text{cov}(1'X\hat{\beta}, 1_s'Y_s) = 1'XA_s^{-1}X_s'1_s = 1'Xc = 1'V1$.

Lemma 1 shows that for models with $V1 \in \mathfrak{M}(X)$, the sample affects the variance only through A_s^{-1} . This simplifies both the study of how the variance depends on the sample and the search for efficient samples.

The collection of samples that satisfy

$$1_s'W_s^{-1/2}X_s/n = 1'X/1'W^{1/2}1,$$

where W is an $N \times N$ matrix, will be denoted by $B(X: W)$. When W is the identity matrix, I , $B(X: I)$ is the collection of samples that are balanced on the columns of X . Royall and Herson (1973) proved that BLU estimators under a wide family of polynomial regression models are greatly simplified in balanced samples:

Theorem 1. Under $M(X: V)$ with $V1 \in \mathfrak{M}(X)$, if $s \in B(X: I)$ then

$$\begin{aligned}\hat{T}(X: V) &= (N/n)1'_s Y_s \\ \text{var}(\hat{T}(X: V)) &= [(N/n) - 1]1' V 1 \sigma^2.\end{aligned}\tag{1}$$

The next theorem shows that if $V = I$ then the variance in (1) is the minimum possible, *i.e.* balanced samples $B(X: I)$, are optimal if $I1 \in \mathfrak{M}(X)$; it also identifies optimal samples for a class of models with more general variance structure.

Theorem 2. Under $M(X: V)$ if both $V1$ and $V^{1/2}1 \in \mathfrak{M}(X)$, then

$$\text{var}(\hat{T}(X: V) - T) \geq [(1' V^{1/2} 1)^2 / n - 1' V 1] \sigma^2;$$

the bound is achieved if and only if $s \in B(X: V)$, in which case

$$\hat{T}(X: V) = (1' V^{1/2} 1) (1'_s V_s^{-1/2} Y_s) / n.$$

Proof: Since $V1 \in \mathfrak{M}(X)$, the quantity to be minimized is $a' A_s^{-1} a$, where $a = X' 1$ (Lemma 1). Now $V^{1/2} 1 \in \mathfrak{M}(X)$ implies that there is a p -vector c_1 for which $V^{1/2} 1 = X c_1$ and, since V is diagonal, this ensures that $V_s^{1/2} 1_s = X_s c_1$ for every sample s . From this it follows that $c'_1 A_s c_1 = n$, and the desired inequality then follows from Schwarz's:

$$(a' A_s^{-1} a) (c'_1 A_s c_1) = (a' A_s^{-1} a) \cdot n \geq (a' c_1)^2.$$

The necessary and sufficient condition for equality is $a' = k c'_1 A_s$, where $k = 1' V^{1/2} 1 / n$. This is equivalent to $s \in B(X: V)$ because $c'_1 A_s = 1'_s V_s^{-1/2} X_s$. The simple forms for the estimator $\hat{T}(X: V)$ and its variance are then easily obtained algebraically.

The formulas in Theorem 2 are familiar in conventional (randomization-based) sampling theory. The BLU estimator $\hat{T}(X: V)$ takes the simple form of the Horvitz-Thompson estimator $\hat{T}_{HT} = \sum_s y_i / \pi_i$, when π_i , the inclusion probability for unit i , is proportional to $v_i^{1/2}$. And the variance bound is the one established by Godambe and Joshi (1965, Theorem 6.1) for the model-based expectation of the random sampling variance.

Suppose that we have, for a working model $M(X: V)$ that satisfies the conditions of Theorem 2, an optimal sample s and BLU estimator \hat{T} . If we now consider a more general model $M(X, Z: V)$ with additional regressor(s) Z , the results of Theorem 2 continue to apply so long as the sample belongs to $B(Z: V)$ as well as to $B(X: V)$. Our sample and estimator remain optimal under the more general model, and the variance is unchanged. That is, we can maintain optimality under our working model (minimum variance sample and BLU estimator) and also protect against bias caused by the additional regressor(s) Z by imposing the additional constraint $B(Z: V)$ on the sample. This procedure not only protects our estimator from bias under $M(X, Z: V)$, it ensures that our sample and estimator both remain **optimal** under the more general model. Of course unbiasedness is ensured under the even more general model $M(X, Z: W)$, where W is any covariance matrix.

3. EXAMPLES

Four models have been particularly prominent in finite population sampling theory. In the polynomial regression model notation of section 1 these are $M(1: 1)$, $M(1, 1: 1)$, $M(0, 1: x)$, and $M(0, 1: x^2)$. Optimal estimators under the first three models are the expansion, regression and ratio estimators, respectively. The optimal estimator under the fourth model,

$\hat{T}(0, 1 : x^2) = \sum_s y_i + (N - n)\bar{x}_r \sum_s (y_i/nx_i)$, is approximated by the mean-of-ratios estimator $\hat{T}_{HT} = N\bar{x} \sum_s (y_i/nx_i)$ when the sampling fraction n/N is small.

One approach to finding a practical sampling and estimation strategy under one of these four working models is to use the best linear unbiased estimator under the model, while ensuring robustness by choosing a sample in which the estimator remains unbiased under more general polynomial regression models. For the first two models, $M(1 : 1)$ and $M(1, 1 : 1)$, we have seen that this strategy produces bias-robustness for free, at no cost in efficiency under the working model. Under both of these models bias protection requires simple (unweighted) balance; but the models satisfy the conditions of Theorem 2 with $V = I$, which implies that simple balance is optimal.

For the other two models, however, there is tension between robustness and efficiency. In section 1 we noted that under $M(0, 1 : x)$ the ratio estimator is optimal, and while the optimal sample consists of the n units maximizing \bar{x}_s , protection from bias under $M(1, 1 : x)$ requires a sample where \bar{x}_s is not maximized but set equal to the population mean, \bar{x} . The situation under $M(0, 1 : x^2)$ is similar: the optimal sample is again the one where the sample mean \bar{x}_s is maximized, but protection of the optimal estimator against bias under polynomial regression models requires an "overbalanced" sample, in which the sample mean equals $\sum_r x_i^2 / \sum_r x_i$ (Scott, Brewer and Ho 1978).

Under both of these models, $M(0, 1 : x)$ and $M(0, 1 : x^2)$, robustness can be achieved at a smaller cost in efficiency by starting with a more general working model. Theorem 2 shows the way. Consider first the model $M(0, 1 : x^2)$. If we use $\hat{T}(0, 1 : x^2)$ in an over-balanced sample, the error variance is $\{(N\bar{x})^2/n - \sum x_i^2 + \sum_s (x_i - \bar{x}_s)^2\}\sigma^2$. But if we use the more general working model $M(0, 1, 1 : x^2)$ and estimator $\hat{T}(0, 1, 1 : x^2)$, the theorem shows that any sample in which $\bar{x}_s = \sum x_i^2 / \sum x_i$ is optimal, yielding the minimum variance $\{(N\bar{x})^2/n - \sum x_i^2\}\sigma^2$. Now bias protection against even more general polynomial regression models can be obtained at no cost in efficiency by imposing the additional constraints of Condition $B(X : V)$ i.e. $\sum_s x_i^{j-1}/n = \sum_1^N x_i^j / \sum_1^N x_i^j$ $j = 0, 3, \dots, J$. Under these constraints on the sample, collectively called π -balance, $T(0, 1, 1 : x^2)$ is the mean-of-ratios estimator (Kott 1984). This sample and estimator remain optimal under all models of the form $M(\delta_0, 1, 1, \delta_3, \dots, \delta_J : x^2)$.

Balanced samples $B(X : V)$ do not always exist. The above example illustrates this; when n becomes so large that $n/N > N(\bar{x}^2) / \sum x_i^2$ there can be no π -balanced sample, because otherwise the variance formula would become negative. Note that the condition $n/N > N(\bar{x}^2) / \sum x_i^2$ implies that $\max(x_i) > N\bar{x}/n$, so that in such populations there is no probability sampling plan with inclusion probability proportional to x .

To generalize the other model, $M(0.1 : x)$, so that the theorem will apply we can add a regressor, $x^{1/2}$:

$$E(Y_i) = \beta_{1/2} x_i^{1/2} + \beta_1 x_i$$

$$\text{var}(Y_i) = \sigma^2 x_i.$$

According to Theorem 2 any sample satisfying

$$\sum_s x_i^{1/2} / n = \sum_1^N x_i / \sum_1^N x_i^{1/2} \quad (2)$$

is optimal under this model, yielding the best linear unbiased estimator $\sum x_i^{1/2} \sum_s x_i^{-1/2} y_i / n$ and the minimum variance, $\{(\sum x_i^{1/2})^2/n - N\bar{x}\}\sigma^2$. This variance compares favorably with

that of the ratio estimator in a balanced sample, $N\bar{x}(N/n - 1)\sigma^2$. Now optimality of the sample and the estimator if in fact $E(Y_i) = \beta_0 + \beta_{1/2}x_i^{1/2} + \beta_1x_i + \beta_2x_i^2$ can be maintained (with no increase in variance) by imposing the additional conditions on the sample:

$$\begin{aligned}\sum_s x_i^{-1/2} / n &= N / \sum_1 x_i^{1/2} \\ \sum_s x_i^{3/2} / n &= \sum_1 x_i^2 / \sum_1 x_i^{1/2}.\end{aligned}\tag{3}$$

These conditions, (2) and (3), give the BLU estimator the simple form:

$$\sum_1 x_i^{1/2} \sum_s (y_i/x_i^{1/2}) / n,$$

which is of course the Horvitz-Thompson estimator for a probability-proportional-to- $x^{1/2}$ sampling plan.

4. PROBABILITY SAMPLING

The results in Section 2 are important in relation to an unobserved regressor Z . If Z were, like X , known for all population units, then we could use $M(X, Z : V)$ as the working model and $\hat{T}(X, Z : V)$ as the estimator in the first place. But suppose that we are unaware of the importance of Z and are using the working model $M(X : V)$ and the estimator $\hat{T}(X : V)$ when in fact $M(X, Z : V)$ applies. In this context we will refer to a sample from $B(X : V)$ as "balanced on X ." Although we can choose a sample that is balanced on X , we cannot ensure that it will be balanced on Z , and if it is not, then our estimator is biased:

$$E(\hat{T}(X : V) - T) = [(1/n)(1'V^{1/2}1)(1'_sV_s^{-1/2}Z_s) - 1'Z]\gamma,$$

where γ is the Z -coefficient: $EY = X\beta + Z\gamma$.

Random sampling can help to provide protection against biases like this. If we use a probability sampling plan with inclusion probabilities, $\pi_i = nv_i^{1/2}/1'V^{1/2}1$, $i = 1, 2, \dots, N$, then we will have balance on Z in expectation:

$$E_{\pi}1'_sV_s^{-1/2}Z_s/n = 1'Z/1'V^{1/2}1,$$

the subscript π indicating that the expectation is with respect to the random sampling plan, not a prediction model. Furthermore, if our sampling plan is one under which $\text{var}_{\pi}(1'_sV_s^{-1/2}Z_s/n)$ approaches zero as n grows, then the probability that we will draw a sample that is badly unbalanced, say one in which $|1'_sV_s^{-1/2}Z_s/n - 1'Z/1'V^{1/2}1| > \delta$, can be made small by taking a large enough sample, n . That is, probability sampling can provide balance on Z "in probability."

The strength of this result is in its scope—it applies for any matrix Z of regressors whatsoever. In particular it applies for the matrix X of regressors in our working model, as well as for

overlooked regressors. The weakness of course is that it applies to the sample selection process, not to a result of that process. The sample actually drawn will, with predictable frequency, be badly unbalanced on the known regressors X . If balance on X is important in a particular study, it should not be left to chance (This was documented empirically by Royall and Cumberland 1981). Restricted random sampling plans which guarantee that the selected sample will be balanced on X , such as Wallenius's "basket method" (1980), might represent a reasonable compromise strategy.

It sometimes happens that a regressor Z that is ignored when the sample is selected becomes available afterwards, as in the case of post-stratification for example. If it is determined that the selected sample is badly balanced on Z , then probability sampling has failed to provide the expected protection against bias under $M(X, Z : V)$; if it is too late to draw another sample, then to protect against the bias we must use an estimator that is unbiased under this model. That is, probability sampling does not guarantee approximate balance on Z ; it only ensures that we have a good chance at approximate balance. It justifies confidence that a given sample is reasonably well balanced, in the absence of evidence to the contrary. It does not justify ignoring evidence of imbalance when it occurs.

Note that under the above probability sampling plan the estimator $(1' V^{1/2} 1)(1_s' V_s^{-1/2} Y_s)/n$, which is $\hat{T}(X : V)$ if both $V1$ and $V^{1/2} 1$ belong to $\mathfrak{N}(X)$ and s is in $B(X : V)$, is unbiased with respect to the probability distribution generated by the sampling plan. But if the sample actually selected is not balanced on X (i.e. if s is not in $B(X : V)$) then this estimator is not unbiased under $M(X : V)$.

REFERENCES

- GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations – I. *Annals of Mathematical Statistics*, 36, 1707-1723.
- ISAKI, C.T., and FULLER, W.A. (1987). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTT, P.S. (1984). A fresh look at bias-robust estimation in a finite population. In *Proceedings of the Section Survey Research Methods, American Statistical Association*, 176-178.
- PEREIRA, C.A., and RODRIGUES, J. (1983). Robust linear prediction in finite populations. *International Statistical Review*, 51, 293-300.
- ROYALL, R.M., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 73, 66-77.
- SCOTT, A.J., BREWER, K.R.W., and HO, W.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73, 359-361.
- TALLIS, G.W. (1986). On the optimality of balanced sampling. *Statistics and Probability*, 4, 141-144.
- TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.
- WALLENIUS, K.T. (1980). Statistical methods in sole source contract negotiation. *Journal of Undergraduate Mathematics and Applications*, 0, 35-47.

Robust Model-Based Methods for Analytic Surveys

T.M.F. SMITH and E. NJENGA¹

ABSTRACT

This paper reviews the idea of robustness for randomisation and model-based inference for descriptive and analytic surveys. The lack of robustness for model-based procedures can be partially overcome by careful design. In this paper a robust model-based approach to analysis is proposed based on smoothing methods.

KEY WORDS: Analytic surveys; Robustness; Smoothing methods.

1. INTRODUCTION

The concept of robustness in finite population inference from both the randomisation and model-based viewpoints is examined. In his seminal paper on a unified theory of sampling from finite populations Godambe (1955) not only proved his famous non-existence theorem but also made suggestions for robust finite population inference. He proposed a superpopulation model for the unit variables y_i and suggested that strategies, that is the choice of both design and estimator, should be based on the model expectation of the sampling variance. He then imposed p -unbiasedness to obtain optimum strategies. These ideas were amplified in several papers including Godambe (1982) and Godambe and Thompson (1977). The results obtained include the apparent optimality of πps sampling and the Horvitz-Thompson (1952) estimator. But the inefficiency of this strategy in multipurpose surveys is well known so we find these results on optimality and robustness less convincing than the apparently negative results on the foundations of inference.

The lack of robustness of many model-based procedures is well known, see Hansen *et al.* (1983), and much of the work of Royall and his colleagues, for example Royall and Herson (1973a,b) has been devoted to constructing robust model-based strategies. After reviewing this work we propose a robust model-based method for estimating many complex statistics employed in the multivariate analysis of survey data which adjusts for the effects of selection. Our proposal is not a strategy but is a procedure which can be employed for the analysis of survey data after the sample is drawn.

2. FORMAL STRUCTURE

In order to examine robustness we must first structure finite population inference in the formal manner pioneered by Godambe (1955). We consider a population of N units with label set $U = \{1, 2, \dots, N\}$. Attached to unit i is a vector of values, y_i , which will be measured on the sample units, and $y_U = (y_1, \dots, y_N)$ denotes the finite population matrix of values. A sample, s , is a subset of U drawn according to some rule. We are concerned here with rules based only on prior information, z_i , available on all the units in the population. Let z_U denote the prior information for the whole population, and let $p(s \mid z_U)$ denote the sampling rule.

¹ T.M.F. Smith, University of Southampton, United Kingdom; E. Njenga, Kenyatta University, Kenya.

Since the rule does not depend on y_U it is uninformative. If $p(s | z_U)$ is a random sampling rule then it determines a probability distribution over ζ , the set of all samples, which is the basis for randomisation inference. The sample data comprises $d_s = \{(i, y_i): i \in s\}$. Let y_s denote the matrix of sample values, then an estimator is a function of the data, d_s , and of the prior information, z_U , which includes auxiliary information. We denote by E_p , V_p , expectations and variances with respect to the distribution $p(s | z_U)$.

In a model-based approach it is further assumed that the population values y_U are random variables. A major problem with this approach is to specify a parametric probability model for the joint distribution of all these random variables, which must be based on all the prior information including that on the structures of, and relationships between, the units in the population. So models must reflect hierarchical groupings (clusters) and block groupings (strata), as well as correlations between the variables. This structure is potentially so complex that attention is usually restricted to means and covariance matrices. In general let $f(y_U | z_U; \lambda)$ denote the conditional finite population distribution, where λ is a vector of unknown parameters. For predictive inference about finite population values, such as totals, this is a sufficient specification. For analytic inference about parameters in the marginal distribution of y we must additionally specify the marginal distribution of the prior values z_U . Let $f(z_U; \phi)$ denote this distribution, then the marginal distribution of y_U is

$$f(y_U; \theta) = \int f(y_U | z_U; \lambda) f(z_U; \phi) dz_U, \quad (2.1)$$

where $\theta = g(\lambda, \phi)$ is the parameter of analytic interest.

Applying the sampling rule to the population generates the data, d_s . The joint distribution of the data, d_s , and prior values, z_U , is

$$\begin{aligned} f(d_s, z_U; \lambda, \phi) &= p(s | z_U) \int f(y_U | z_U; \lambda) f(z_U; \phi) dy_{\bar{s}} \\ &= p(s | z_U) f(y_s | z_U; \lambda) f(z_U; \phi), \end{aligned} \quad (2.2)$$

where \bar{s} denotes units not in s . This distribution is the basis of a model-based approach to inference. We let E_m , V_m , denote expectations and variances with respect to the model.

An implication of (2.2) is that the sampling rule, $p(s | z_U)$, must be completely known to the person making the inference, as must the values of z_U . Absence of knowledge may render $p(s | z_U)$ informative about the unobserved values $y_{\bar{s}}$, see Scott (1977), Sugden and Smith (1984), in which case it cannot be taken outside the integral in (2.2).

In this general set-up, embracing both random selection and modelling of values, randomisation inference corresponds to the case where the values y_U are unknown constants and the model distribution becomes degenerate at the point y_U . The only probability remaining is that in $p(s | z_U)$, and this distribution over the set ζ of all possible samples is the basis of randomisation inference. Note that the randomisation distribution is completely specified by knowledge of the sampling rule and of the prior values, z_U . It does not depend on any unknown parameters or on the survey values, y_U . This renders $p(s | z_U)$ uninformative because there is less information in $p(s | z_U)$ than in z_U itself. This accounts for the negative nature of Godambe's results about randomisation inference.

In contrast model-based inference depends solely on the model component of (2.2), since $p(s | z_U)$ contains no information about $y_{\bar{s}}$. Predictive inferences about $y_{\bar{s}}$ are made using the conditional distribution, $f(y_{\bar{s}} | y_s, z_U; \lambda)$, independent of the randomisation distribution, $p(s | z_U)$. The sampling rule is still important at the design stage, for it affects efficiency and robustness, but it has no rôle to play at the inference stage. Random sampling also provides

a guarantee that the sampling rule is in fact uninformative, providing a scientifically acceptable sampling procedure. Model-based inferences may not be robust, however, because they may depend strongly on the choice of model, as demonstrated by many authors including Hansen *et al.* (1983).

A compromise solution is to employ both components of (2.2), the model and the randomisation distribution, in the choice of estimator. This was proposed by Godambe (1955) as a positive response to his negative results. He proposed using as a criterion the model expectation of the randomisation variance, namely $E_m V_p(t_s)$, where t_s is an estimator of a finite population total T . To find an optimum solution in a particular class of models Godambe restricted the choice of t_s to the class of p -unbiased estimators. This restriction has been much criticized and subsequently several authors, including Brewer (1979), Särndal (1980), Isaki and Fuller (1982), Little (1983), have proposed replacing exact unbiasednesses by some form of approximate unbiasedness. This is usually expressed in the form of asymptotic design unbiasedness which requires the construction of a hypothetical sequence of finite populations with sizes tending to infinity. Although one may feel unhappy with this mathematical construction the suggestion that strategies, chosen before drawing the sample, should be based on considerations of the average under a model of a repeated sampling procedure is perfectly acceptable. The controversial issue is the choice of distribution for making inferences after the sample has been drawn.

3. ROBUSTNESS

Robustness is not a well defined concept in statistics. The Encyclopedia of Statistical Sciences, (Kotz and Johnson 1988), states that:

“a robust procedure performs well not only under ideal conditions but also under departures from the ideal.”

It goes on to say that both the nature of departures from the ideal and the meaning of “*performs well*” must be specified. With this broad definition in mind we now examine robustness for randomisation and model-based inference for finite population totals. The general perception is that randomisation inference is robust and that model-based inference is not.

Godambe’s negative results can be interpreted to mean that randomisation inference is impossible in general. This is certainly true for heterogeneous populations, such as Royall’s axe, ass and box of horseshoes, or for populations with a few very extreme values, but for homogeneous populations the evidence overwhelmingly shows that randomisation inference is not only possible but also works in a well defined sense.

Employing randomisation inference implies abandoning certain statistical principles, such as the likelihood principle, and replacing them by an appeal to the central limit theorem. The assertion is that under repeated random sampling using the specified rule $p(s \mid \mathcal{Z}_U)$

$$\frac{t_s - T}{\hat{V}_p(t_s)} \sim N(0,1), \quad (3.1)$$

for any t_s which is approximately p -unbiased for T , where both N and n are large, but n/N is small. Although proved formally only under SRS and related schemes, empirical evidence shows that the randomisation coverage properties of 95% confidence intervals of the form

$$t_s \pm 1.96\sqrt{\hat{V}_p(t_s)}, \quad (3.2)$$

where $\hat{V}_p(t_s)$ is a consistent estimator of $V_p(t_s)$, are approximately correct except for extreme designs or heterogeneous populations.

Godambe and Thompson (1977) express their views about this approach in the following terms.

“The use of such a confidence interval may be interpreted as follows:

I: We are fairly sure a priori that y belongs to that subset of R^N for which the interval covers $T(y)$ for 95% of all possible samples.

II: There is no way that the sampled y -values, in conjunction with whatever other information we may have about the population, have altered the conviction in I. Thus even after sampling we believe that if the design were implemented again and again on this population the interval would cover $T(y)$ approximately 95% of the time.

The robustness of the interval arises of course from the fact that only very weak and essentially informal conditions are required for the validity of its interpretation in the sense of I and II.”

Very similar views are expressed by Hansen *et al.* (1983).

“For probability-sampling designs the computed confidence intervals, for samples large enough, are valid in the sense that the randomization probability that the confidence intervals contain the value being estimated is equal to or greater than the nominal confidence coefficient, independent of the distribution of the characteristics among the elements of the population from which the sample is drawn.”

“Robustness is usually understood to mean that inferences made from a sample are insensitive to violations of the assumptions that have been made. In principle, and ordinarily in fact, robustness is achieved in probability-sampling surveys by the use of sampling with known probabilities (i.e., randomization) and consistent estimators, and using a large enough sample that the central limit theorem applies, so that the estimates can be regarded as approximately normally distributed.”

Note that this concept of robustness does not appear to require any specification of ideal conditions or of departures from the ideal. Random sampling and consistent estimation are all that is required. Brewer and Särndal (1983) are quite explicit:

“Probability sampling methods are robust by definition; since they do not appeal to a model, there is no need to discuss what happens under model breakdown.”

How can a statistical procedure be so robust?

The reason is that the entire procedure is under the control of the statistician, no attempt is made to introduce “nature” into the structure. The randomisation distribution has a known form and does not depend on unknown parameters. There is no need to make an inference about $p(s \mid z_U)$. Similarly the framework for inference is chosen by the statistician, it is repeated sampling using $p(s \mid z_U)$. Different statisticians may use different sampling rules and estimators but the procedure represented by (3.1) gives approximately correct coverage properties in every case, and so is robust. This is an example of criterion robustness. However, any given procedure may not be efficient for the totals of some variables. We have already highlighted the well known inefficiency of the Horvitz-Thompson estimator which occurs when

the survey variable is negatively correlated with the size variable. The search for efficiency robustness over a wide range of variables leads frequently to the recommendation that the design should be a stratified SRS design, see, for example, Godambe (1982), Hansen *et al.* (1983).

In model-based inference the statistician is playing the game of modelling "nature". Probability distributions such as $f(y_U | z_U; \lambda)$ are chosen by the statistician but their true form is unknown, as also are the values of the parameters. If an estimator, t_s of T , is chosen then its expected value and variance will depend on the choice of model. Deviations from the model may lead to changes in the mean and variance and hence to changes in confidence intervals based on applying the central limit theorem to the model residuals. In model-based inference the robustness due to the central limit theorem is more limited than that in randomisation inference since it applies only to the residuals. Some model deviations can be controlled by choosing an appropriate design, as in Royall and Herson (1973a,b), but there can never be complete robustness. The framework for inference is also completely different. Instead of employing the unconditional distribution based on repeated sampling model-based inference employs the conditional distribution given the selected sample s .

Can these two positions ever be reconciled? Before sampling, when choosing strategies, they can. Both schools of thought have the same prior information, z_U , and both use models to suggest designs and estimators and choose strategies based on the overall mean squared error

$$E_m E_p (t_s - T)^2. \quad (3.3)$$

Randomisers usually impose a constraint such as approximate p -unbiasedness while modellers may impose approximate model unbiasedness and the two positions can be reconciled by choosing a sample design such that the model-unbiased estimator is also p -unbiased. This strategy utilizes the full structure of (2.2) and gets the best of both worlds.

After sampling there appears to be little hope of reconciliation. The two frameworks for inference are quite different, one being based on an unconditional distribution the other on a conditional distribution. Royall and Cumberland (1981) have demonstrated convincingly how much difference this can make. Incidentally they have also demonstrated the lack of robustness of some of the conventional model-based variance estimators.

One case where reconciliation is possible occurs in stratified sampling. Both randomisers and modellers have converged on stratified sampling as a robust design, and for SRS within strata model-based and p -based inferences coincide. This provides evidence for one of the few positive results in sample surveys:

Theorem: Stratification is a good thing.

Proof: See Cochran (1977, Ch.5).

Stratification allows us to look at the problem of robustness more closely. If both a randomiser and a modeller adopt the same stratification, and both also adopt the same SRS design within strata, then for a given sample they will both make identical inferences. Now suppose on the basis of further analysis or evidence it is agreed that an extra level of stratification should have been used. How does this affect the respective inferences? The modeller now has to say that the original model was misspecified and hence that inferences from that model would be biased. Both the estimator and the variance of the original model would be wrong. The randomiser, however, can say that the extra information is interesting, and could be used to post-stratify the original results, but that it can also be ignored if necessary because the original inferences are still valid in the sense defined in (3.2). All that has happened is a possible loss of efficiency. In one case the original inference is condemned as not being robust, in the other case the same

inference is apparently robust. The modellers bias, when averaged over repeated samples, is transformed for the randomiser into a component of sampling variance, or a loss of efficiency. So if initially randomisers and modellers start from the same position then deviations from that position are interpreted differently. In one case it is a bias in the other case a variance. Can this really be called robust in one case and not robust in the other?

4. ANALYTIC INFERENCE

In analytic inference the target for inference is no longer a known function of the finite population values, y_U , so that even if $n = N$ there is still residual uncertainty in the inference. Examples are tests of hypotheses, where the null hypothesis of no difference is meaningless in a fixed finite population. Possible targets for inference are the parameters λ, ϕ , of the model (2.2), or functions of them such as θ in (2.1). Other targets are the parameters in finite populations related to the given finite population in some known way, perhaps through a spatial or time series structure. Methods for analytic inference have recently been reviewed by Skinner *et al.* (1989).

The starting point for analytic inference is the specification of the superpopulation model which aims to show how the finite population is related to the superpopulation. A common assumption is that the finite population is generated as IID random variables from a superpopulation. Whether this can be justified for populations with structure, such as clustering or stratification, is debatable. In this paper we assume that it is true, at least within broadly defined strata. With this assumption a SRS from the finite population is itself an IID sample from the superpopulation and inferences can be made directly from the sample to the superpopulation. If the sample is not a SRS, but is drawn using a design $p(s \mid z_U)$ which uses the information in z_U , then the achieved sample is no longer an IID sample from the superpopulation. This is the problem of selection and the effect of selection must be taken into account in the final inference.

The superpopulation model establishes a hierarchy,

$$\text{superpopulation} \supset \text{finite population} \supset \text{sample}.$$

If the finite population is IID from the superpopulation then finite population parameters, such as means, are related to the corresponding superpopulation parameters by

$$\bar{y}_U = E_m(\bar{y}_U) + O_p(N^{-1/2}). \tag{4.1}$$

Since N is usually very large an inference about \bar{y}_U is a good approximation to an inference about $E_m(\bar{y}_U)$. Inferences about \bar{y}_U using the p -weights associated with the sampling rule $p(s \mid z_U)$ are the basis of the randomisation approach to analytic inference. Note that this approach depends strongly on the IID assumption for the finite population.

For more complex analyses, such as logistic regression analysis, the pseudo-MLE approach in Skinner *et al.* (1989, sec. 3.4.4.) and Binder (1983) can be used to define both the finite population parameter of interest and the randomisation estimator. The finite population parameter is usually defined through an estimating equation, see Godambe (1960) and Godambe and Thompson (1986). As in Section 3 confidence intervals are based on the unconditional distribution generated by repeated random sampling.

Model-based analytic inference is based on the complete model of the survey population y_U , the design variables z_U , and the sample selection rule $p(s \mid z_U)$, that is

$$f(\chi_U, z_U, s; \lambda, \phi) = f(\chi_U | z_U; \lambda) f(z_U; \phi) p(s | z_U). \quad (4.2)$$

For random sampling rules the selection scheme leaves the conditional distribution $f(\chi_U | z_U; \lambda)$ unchanged, but changes the marginal distribution of z_U from $f(z_U; \phi)$ before selection to

$$g_s(z_U; \phi) = f(z_U; \phi) p(s | z_U) \quad (4.3)$$

after selection. Thus inferences about λ are unaffected by selection but inferences about ϕ , and hence about $\theta = g(\lambda, \phi)$, the parameters of the marginal distribution $f(\chi_U; \theta)$, are affected by selection. For these latter inferences the sample data cannot be treated as though it were a SRS from the superpopulation model.

If we assume that the superpopulation distributions are multivariate normal then

- (i) $E(\chi | z)$ is linear in z , and
- (ii) $V(\chi | z) = K$, independent of z .

Under these assumptions of linearity and homoscedasticity a model-based estimator of the covariance matrix, Σ_{yy} , of y is given by

$$\hat{\Sigma}_{yy} = Y_{yys} + b_{yz} (V_{zzu} - V_{zzs}) b_{yz}^T, \quad (4.4)$$

as shown in Skinner *et al.* (1989 Section 6.4), where Y_{yys} , V_{zzs} , b_{yz} are sample covariance matrices and a matrix of regression coefficients based on treating the sample data as IID from the conditional distribution $f(\chi_U | z_U; \lambda)$. We call (4.4) the Pearson adjusted estimator after Pearson (1903).

Theoretical and empirical studies by Pfeffermann and Holmes (1985), Holmes (1987) and Njenga (1990), have shown that model-based inferences from (4.4) are not robust to departures from the assumptions of linearity and homoscedasticity. Nathan and Holt (1980) proposed a p -weighted version of (4.4) as a more robust alternative. This estimator is formed by replacing all the equally weighted sums in (4.4) by the corresponding p -weighted sums. The resulting estimator is called the probability weighted maximum likelihood estimator (*pwml*). The properties of this estimator have been studied empirically and theoretically in Holmes (1987), Njenga (1990) and in Skinner, Holt and Smith (1989, Ch.8). It was found to have similar unconditional properties to alternative p -weighted estimators, such as the Horvitz-Thompson estimator of Σ_{yy} , and superior conditional properties. In the simulation study in Section 6 the *pwml* estimator is taken to represent the entire class of p -weighted estimators. Since the p -weighted version of V_{zzs} in (4.4) is a design consistent estimator of V_{zzu} the resulting estimator is a design consistent estimator of Σ_{yy} . We now investigate a new robust model-based procedure.

5. A NONPARAMETRIC MOMENT-BASED ESTIMATOR

In this section we attempt to overcome the lack of robustness of model-based estimators such as (4.4) which depend strongly on assumptions of linearity and homoscedasticity. If the finite population is realized as IID observations from the superpopulation and if interest centres on the superpopulation parameters μ_y, Σ_{yy} in the marginal distribution of y , then the approach we adopt uses the fact that the sample data are IID from the conditional distribution $f(y | z)$

while the design variables \underline{z}_U are an IID sample of size N from the marginal distribution of \underline{z} . For simplicity we assume that only one design variable has been used, such as a measure of size, so that \underline{z} is a scalar random variable.

We assume that the conditional mean and covariance matrix of y given \underline{z} are smooth functions of \underline{z} of unknown form. Let

$$E(y \mid \underline{z}) = \mu(\underline{z}), \quad (5.1)$$

$$V(y \mid \underline{z}) = \Sigma_{yy}(\underline{z}). \quad (5.2)$$

These parametric functions can be estimated using some form of nonparametric estimation such as linear smoothing. Examples of linear smoothing methods are kernel estimation, see, for example, Gasser and Muller (1979), local regression, see, for example, Cleveland (1979), and smoothing splines, see, for example, Silverman (1985). We propose estimating the functions in (5.1) term by term using the kernel estimator

$$\hat{\mu}(\underline{z}) = \sum_{j \in S} W_k(\underline{z}, \underline{z}_j) y_j. \quad (5.3)$$

We constrain the sum of the weights to be unity so that the estimator is a weighted average and employ the Gaussian kernel with k being the bandwidth. These estimators have been extensively studied and a recent review is Gasser and Engel (1990).

The structure in (5.1) and (5.2) implicitly assumes that we can write

$$y_j = \mu(\underline{z}_j) + \varepsilon_j, \quad j \in S, \quad (5.4)$$

so that

$$\hat{\varepsilon}_j = y_j - \hat{\mu}(\underline{z}_j), \quad j \in S. \quad (5.5)$$

Thus

$$\hat{\varepsilon}_j \hat{\varepsilon}_j^T = (y_j - \hat{\mu}(\underline{z}_j))(y_j - \hat{\mu}(\underline{z}_j))^T \quad (5.6)$$

is an estimator of $\Sigma_{yy}(\underline{z}_j)$. Applying a linear smoother to each term $\sigma_{ab}(\underline{z}_j)$ of $\Sigma_{yy}(\underline{z}_j)$ gives

$$\hat{\sigma}_{ab}(\underline{z}) = \sum_{j \in S} W_h(\underline{z}, \underline{z}_j) \hat{\varepsilon}_{ja} \hat{\varepsilon}_{jb}, \quad (5.7)$$

where $W_h(\underline{z}, \underline{z}_j)$ is a kernel with band width h which will usually be wider than the band width k chosen for the estimation of the conditional mean, (5.3).

The estimates of the marginal moments then employ the standard results that

$$\mu_y = E_z(\mu(\underline{z})), \quad (5.8)$$

$$\Sigma_{yy} = E_z(\Sigma_{yy}(\underline{z})) + V_z(\mu(\underline{z})). \quad (5.9)$$

Now

$$\mu_y = \int \mu(z)f(z)dz,$$

and our proposed estimator is

$$\hat{\mu}_y = \int \hat{\mu}(z)\hat{f}(z)dz. \quad (5.10)$$

Since N is large we propose using the empirical p.d.f. (Parzen 1962), given by

$$\begin{aligned} d\hat{F}(z) &= \hat{f}(z) = 1/N, \quad \text{if } z = z_j, \quad j = 1, \dots, N, \\ &= 0, \quad \text{otherwise.} \end{aligned} \quad (5.11)$$

Substituting in (5.10) gives the estimator

$$\hat{\mu}_y = N^{-1} \sum_{j=1}^N \hat{\mu}(z_j). \quad (5.12)$$

To estimate Σ_{yy} we adopt a similar procedure for the first term of (5.9). The second term can be written

$$V_z(\mu(z)) = \int (\mu(z) - \mu_y)(\mu(z) - \mu_y)^T f(z)dz. \quad (5.13)$$

For our estimator we propose

$$\hat{V}_z(\mu(z)) = N^{-1} \sum_{j=1}^N (\hat{\mu}(z_j) - \hat{\mu}_y)(\hat{\mu}(z_j) - \hat{\mu}_y)^T. \quad (5.14)$$

Thus the proposed estimator of is Σ_{yy} is

$$\hat{\Sigma}_{yy} = N^{-1} \left[\sum_{j=1}^N \{\hat{\Sigma}_{yy}(z_j) + (\hat{\mu}(z_j) - \hat{\mu}_y)(\hat{\mu}(z_j) - \hat{\mu}_y)^T\} \right]. \quad (5.15)$$

Njenga (1990) examines the asymptotic statistical properties of these estimators.

One of the main reasons for estimating Σ_{yy} is to carry out some form of multivariate analysis, such as a regression analysis between two or more of the components of y . In the next section we report the results of a simulation study in which the simple regression coefficient between two y -variables is estimated from stratified random samples with different sampling fractions.

6. ESTIMATING A REGRESSION COEFFICIENT A SIMULATION STUDY

Let $y = (y_1, y_2)^T$ with mean $\mu_y = (\mu_1, \mu_2)^T$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

We are interested in estimating a function of \sum_{yy} , the simple linear regression coefficient,

$$B_{12} = \sigma_{12}/\sigma_2^2. \quad (6.1)$$

The elements of \sum_{yy} will be estimated using:

- (i) the Pearson adjusted estimator of \sum_{yy} based on (4.4),
- (ii) the probability weighted version of (4.4),
- (iii) a kernel estimator based on (5.14).

The corresponding estimators of B_{12} , or of its finite population equivalent B_{12U} , are denoted $\hat{B}_{12,ml}$, $\hat{B}_{12,pwml}$ and $\hat{B}_{12,nw}$ respectively. The estimator $\hat{B}_{12,ml}$ is indexed by “ml” because it is also the MLE under a multivariate normal model. The estimator $B_{12,nw}$ is indexed “nw” after Nadaraya (1964) and Watson (1964). The first two estimators were chosen because of their good performance in previous simulation studies, see Skinner *et al.* (1989, Ch.8).

We carried out three types of simulation study. In the first simulation study we generated a multivariate normal population to compare the performance of the new estimator with the maximum likelihood estimator which is optimal for this population. In the second simulation study we generated a quadratic homoscedastic population to compare the estimators when only the linearity assumption is violated. In the last simulation study we compared the estimators when the structure of the population is unknown, *i.e.* we used a ‘real’ population. In these simulation studies we carried out both conditional and unconditional analyses. The former allow us to assess whether a particular estimator is good in some samples and poor for others whereas the latter averages over all possible samples for a particular design.

The new estimator uses the Gaussian Kernel

$$W_k(z_i, z_j) = c_i \exp\{-(z_i - z_j)^2/2k^2\}, \quad i \in U, \quad j \in s,$$

where $c_i = 1/\sum_{j \in s} \exp\{-(z_i - z_j)^2/2k^2\}$. A simulation with different values of the band width k showed that the mean squared error was relatively constant for a wide range of values of k and that this was achieved by trading off bias against variance. We selected values for k that gave relatively small values for the bias for each stratified sample design.

Since the ‘real’ population available to us was 6,962 observations from the 1975 UK Family Expenditure Survey we constructed all three populations to be of this size with mean vector and covariance matrix

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_z \end{bmatrix}, \quad \underline{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1z} \\ & \sigma_2^2 & \sigma_{2z} \\ & & \sigma_z^2 \end{bmatrix}.$$

The actual values of $\underline{\Sigma}$ are shown in Table 6.1.

The design variable is based on the expenditure on food, the independent variable is the total income and the dependent variable is the total expenditure. This finite population was stratified into five strata according to increasing values of the design variable, such that the first stratum contains 1,393 units with lowest values of z , second, third, fourth contain 1,392 units each and the fifth contains the last 1,393 units with the highest z values.

Table 6.1
Parameter Values from the Real Population

Variable		S.D.	Correlation matrix			
y_1	Expenditure on all items	0.668	1			
y_2	Total income	0.849	0.75	1		
z	Expenditure on food	0.658	0.41	0.28	1	

Table 6.2
Stratified Sample Designs

Sample design		n_1	n_2	n_3	n_4	n_5	Symbol
D1	Proportional allocation	20	20	20	20	20	Δ
D2	Increasing allocation	5	9	16	30	40	∇
D3	U-shaped allocation	40	8	4	8	40	+

The sample designs used were based on those used by Holt, Smith and Winter (1980). Denote a stratified random sampling design by $(n_1 \dots n_5)$ with n_h units selected from the h^{th} stratum, $h = 1, \dots, 5$, then the designs are shown in Table 6.2, together with the symbols used in the plots.

For the various stratified sample designs we selected 1,000 independent samples of size 100 from the finite population. The sampling distribution of the various statistics under investigation were estimated from these 1,000 repeated samples. We obtain the unconditional results by averaging the statistics under investigation over all the 1,000 samples.

To assess the conditional properties of the estimators the 1,000 samples were divided into 20 groups of 50 samples each according to increasing values of $\Delta_{zz}^F = (S_{zzs} - S_{zz})/S_{zz}$ for the nw and ml estimators where

$$S_{zz} = N^{-1} \sum_U (z_i - \bar{z}_U)^2, \quad S_{zzs} = n^{-1} \sum_s (z_i - \bar{z}_s)^2,$$
$$\bar{z}_U = N^{-1} \sum_U z_i, \quad \bar{z}_s = n^{-1} \sum_s z_i,$$

and of $\Delta_{zz}^{*F} = (S_{zzs}^* - S_{zz})/S_{zz}$ for the $pwml$ estimators where

$$S_{zzs}^* = \sum_s w_i (z_i - \bar{z}_s^*)^2, \quad \bar{z}_s^* = \sum_s w_i z_i, \quad w_i = (N\pi_i)^{-1} \quad \text{and} \quad \pi_i$$

denotes the probability of including the i^{th} unit in the sample such that the first group contained the 50 samples with the smallest values of Δ_{zz}^F (or Δ_{zz}^{*F}) and so on up to the 20th group which contains the 50 samples with the largest values of Δ_{zz}^F (or Δ_{zz}^{*F}). We assume that the variation in Δ_{zz}^F (or Δ_{zz}^{*F}) within each group is small. The conditional distribution of the various estimators given Δ_{zz}^F (or Δ_{zz}^{*F}) can then be plotted.

The biases, standard deviations and mean square errors reported in simulation studies 1 and 2 are computed around the value of B_{12U} in the finite population generated from the model. This enables them to be compared with the values generated from the real finite population in simulation study 3.

Table 6.3
Unconditional Absolute Biases of the Three Estimators of B_{12}
 $N = 6,962, n = 100$ True Value $B_{12} = 0.595$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0003	0.0003	0.0185
D2	0.0007	0.0019	0.0269
D3	0.0026	0.0018	0.0159

Table 6.4
Unconditional Standard Deviation of the Three Estimators of B_{12}

Sample design	Standard deviations		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0500	0.0500	0.0507
D2	0.0522	0.0693	0.0531
D3	0.0486	0.0710	0.0503

Table 6.5
Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0025	0.0025	0.0029
D2	0.0027	0.0048	0.0035
D3	0.0024	0.0050	0.0028

Simulation Study 1

In the first simulation study the 6,962 finite population values were generated from a multivariate normal distribution with correlation matrix given in Table 6.1. These data should be favourable to the estimator $\hat{B}_{12,ml}$.

The unconditional biases, standard deviations and mean squared errors are shown in Tables 6.3, 6.4 and 6.5.

As expected the estimator $\hat{B}_{12,ml}$ is best in terms of mean squared error. The new estimator $\hat{B}_{12,nw}$ does surprisingly well, it has a large bias but a similar standard deviation. The size of the bias for a very smooth (linear) population is consistent with the results in other studies, see Gasser and Engel (1990). A very wide bandwidth is needed to capture a very smooth function.

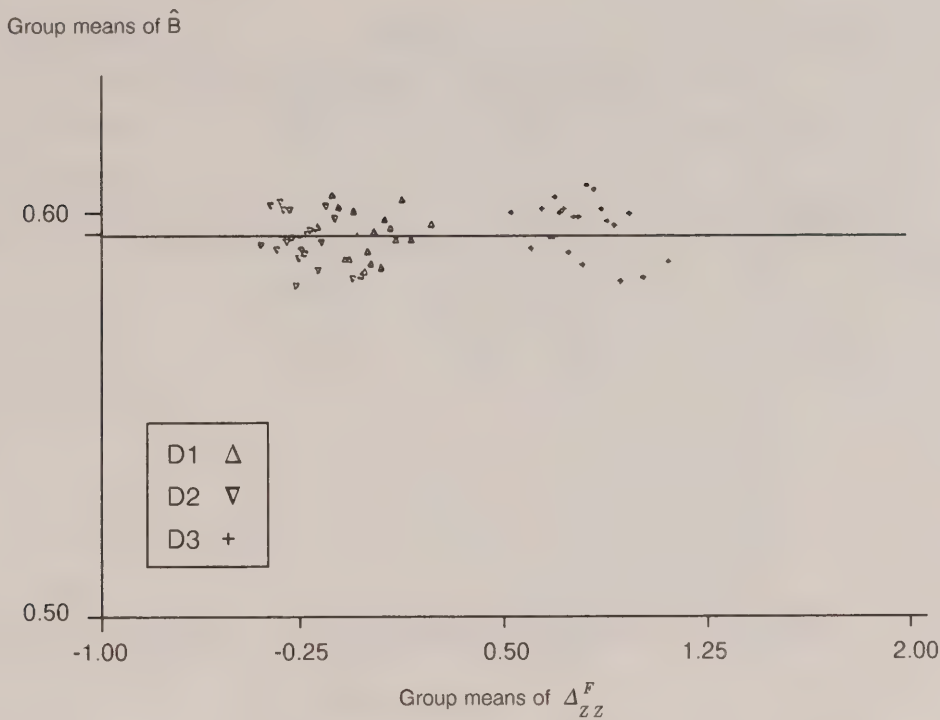


Figure 6.1 Scattergram of group means of $\hat{B}_{12,ml}$

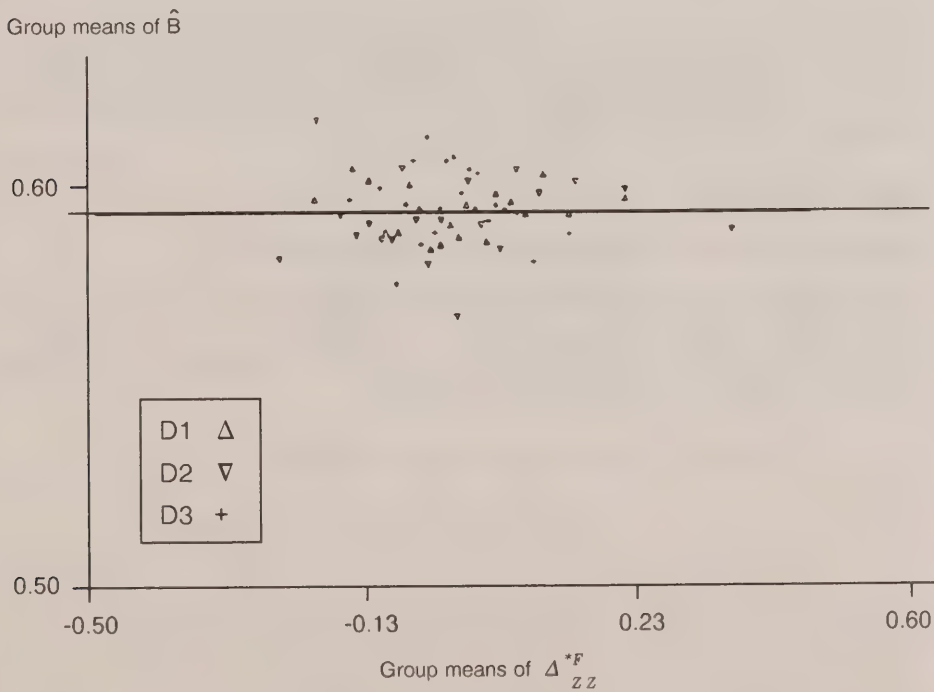


Figure 6.2 Scattergram of group means of $\hat{B}_{12,pwml}$

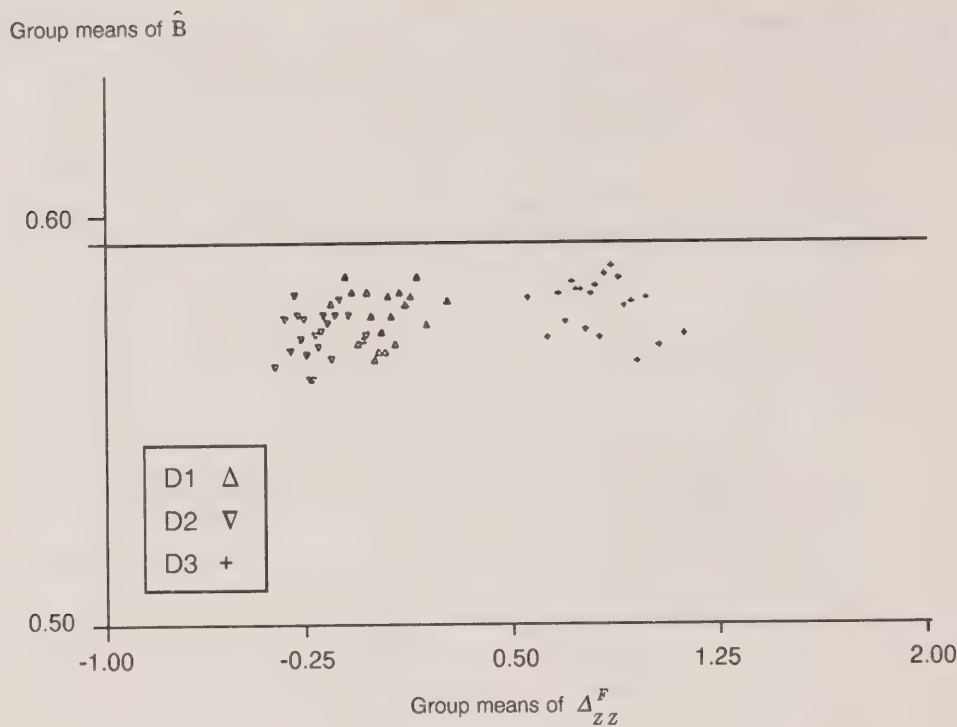


Figure 6.3 Scattergram of group means of $\hat{B}_{12,nw}$

The conditional plots are shown in Figures 6.1, 6.2 and 6.3. These plots show that there is no additional pattern to the bias beyond the absolute level of bias shown in Table 6.3. Previous studies have shown consistent patterns of bias for SRS estimators and simple p -weighted estimators, see Skinner *et al.* (1989, Chs. 7 and 8).

Simulation Study 2

Repeated sampling from a quadratic homoscedastic population

This simulation study is similar to one carried out by Holmes (1987). We generated 6,962 finite population values of (y_{1i}, y_{2i}, z_i) $i = 1 \dots 6,962$ by first generating a value of z_i from the uniform distribution $U(0,10)$. Using this generated value of z_i the corresponding values of y_{1i} and y_{2i} are obtained from the relationships;

$$y_{2i} = m_2 + H_2 z_i + R_2 z_i^2 + \epsilon_{2i}$$

and

$$y_{1i} = m_1 + H_1 z_i + R_1 z_i^2 + \epsilon_{1i},$$

where ϵ_{2i} and ϵ_{1i} are random variables from normal distributions with mean zero and constant variance, and $R_1 \neq 0$, $R_2 \neq 0$. Following Holmes (1987) we chose the parameters in these expressions so that the regressions of y_1 and y_2 on z are monotonically increasing functions of z and the regression of y_1 on y_2 is approximately linear so that the regression coefficient B_{12} will be a meaningful parameter to estimate.

Table 6.6
Unconditional Standard Deviation of the Three Estimators of B_{12}
 $N = 6,962, n = 100$ True Value $B_{12} = 0.857$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0119	0.0119	0.0171
D2	0.0923	0.0132	0.5556
D3	0.0124	0.0098	0.0104

Table 6.7
Unconditional Standard Deviation of the Three Estimators of B_{12}

Design	Standard deviations		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0877	0.0877	0.0877
D2	0.0972	0.1230	0.1150
D3	0.0785	0.1110	0.0797

Table 6.8
Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0078	0.0078	0.0080
D2	0.0180	0.0153	0.0164
D3	0.0063	0.0124	0.0065

The unconditional results of the three estimators of the regression coefficient are given in Tables 6.6, 6.7 and 6.8.

We see from the tables that the *ml* estimator is severely biased and very inefficient for the increasing allocation design D2, but is approximately unconditionally unbiased and efficient for the designs D1 and D3. The *pwml* estimator as expected is approximately unconditionally unbiased across all the sample designs considered. Though more biased than the *pwml* estimator, the *nw* estimator is less biased than the *ml* estimator for the unequal probability designs. We also see that the *nw* estimator is more efficient than *ml* for the design D2 and approximately equally efficient for design D3. It is also more efficient than the *pwml* estimator for the U-shaped design D3.

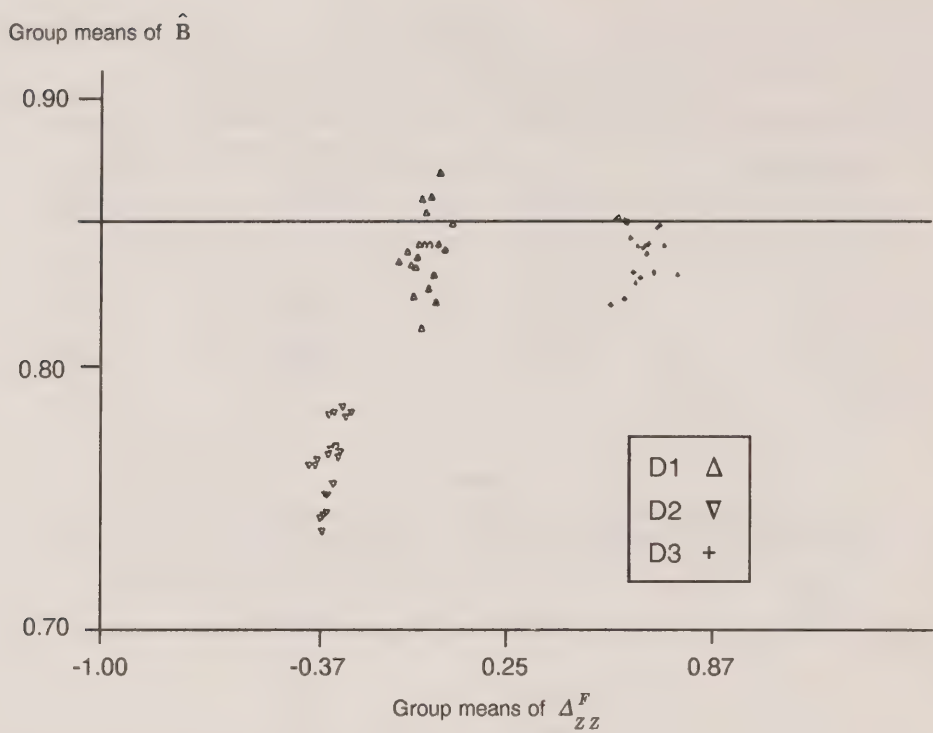


Figure 6.4 Scattergram of group means of $\hat{B}_{12,ml}$

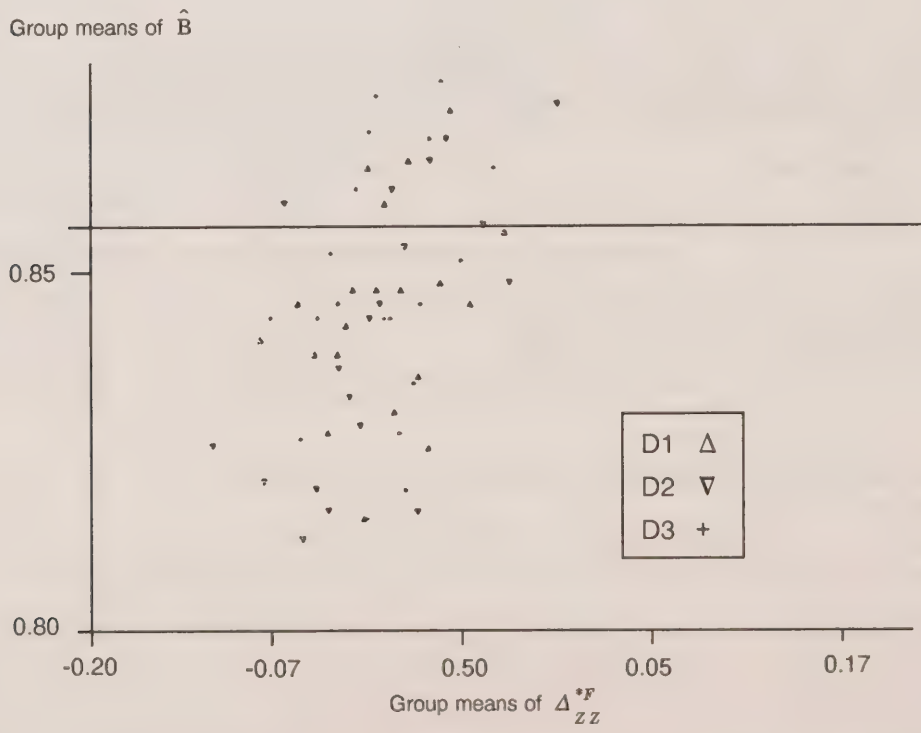


Figure 6.5 Scattergram of group means of $\hat{B}_{12,pwml}$

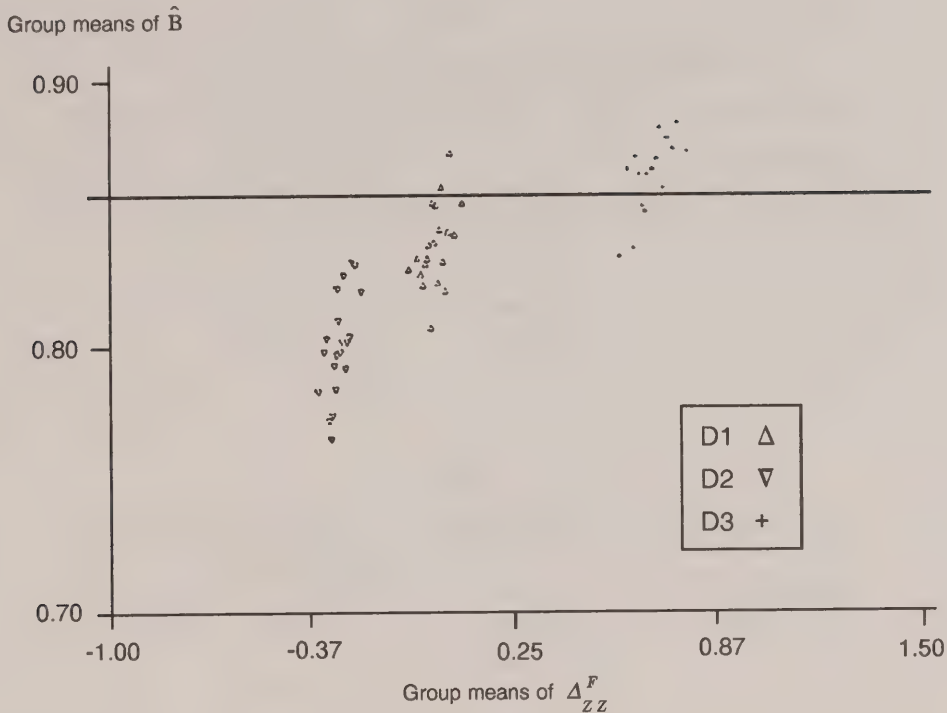


Figure 6.6 Scattergram of group means of $\hat{B}_{12,nw}$

The plots of the conditional analysis are shown in Figures 6.4, 6.5 and 6.6.

We see from Figure 6.4 that the *ml* estimator is approximately conditionally unbiased for the design D1 and D3, and has no additional conditional bias for the design D2. From Figure 6.5 we see that the *pwml* estimator has no additional conditional bias for any of the designs. We see from Figure 6.6 that the *nw* kernel estimator has only a small additional conditional bias within each of the three probability designs.

Simulation Study 3

Repeated sampling from a multivariate ‘Real’ population

In this simulation study we employ the 6,962 actual data points from the Family Expenditure Survey for the finite population. We consider the same variables as in section 3.1 and sample repeatedly from this population to investigate the robustness properties of the three regression estimators. We expect the real population to violate all the normality assumptions.

The unconditional results are shown in Tables 6.9, 6.10 and 6.11, and we see that the *nw* kernel estimator is the most efficient and is approximately unconditionally unbiased across all the probability designs. The *ml* estimator is less biased and more efficient than the *pwml* estimator for the unequal probability designs.

The plots of the conditional analyses are shown in Figures 6.7, 6.8 and 6.9.

We see from Figure 6.7 that the *ml* estimator is approximately conditionally unbiased for the designs D1 and D2 but has a slight conditional bias for design D3. From Figure 6.8 we see that the *pwml* estimator has no additional conditional bias for any of the designs. From Figure 6.9 we see that the *nw* kernel estimator is approximately conditionally unbiased for the three probability designs.

Table 6.9
Unconditional Absolute Biases of the Three Estimators of B_{12}
 $N = 6,962, n = 100$ True Value $B_{12} = 0.595$

Sample design	Absolute biases of		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0245	0.0245	0.0056
D2	0.0260	0.0408	0.0060
D3	0.0128	0.0355	0.0072

Table 6.10
Unconditional Standard Deviation of the Three Estimators of B_{12}

Sample design	Standard deviation		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.111	0.111	0.111
D2	0.106	0.132	0.108
D3	0.111	0.122	0.111

Table 6.11
Unconditional Mean Square Errors of the Three Estimators of B_{12}

Sample design	Mean square errors		
	$\hat{B}_{12,ml}$	$\hat{B}_{12,pwml}$	$\hat{B}_{12,nw}$
D1	0.0130	0.0130	0.0121
D2	0.0120	0.0192	0.0117
D3	0.0125	0.0161	0.0123

We conclude from these simulation studies that the new estimator $\hat{\beta}_{12,nw}$ has performed well. When the assumptions of linearity and homoscedasticity are violated it appears to be robust across a variety of designs, to have good efficiency and to have reasonable conditional as well as unconditional properties. We know from previous studies that $\hat{\beta}_{12,pwml}$ performs as well as more conventional p -weighted estimators unconditionally and has far better conditional properties. The fact that in this study the new estimator $\hat{B}_{12,nw}$ apparently has better properties than the $pwml$ estimator, which was chosen to represent the class of p -weighted estimators because of its performance in other simulation studies, suggests that it is an approach that could be considered in analytic studies of a small number of key parameters.

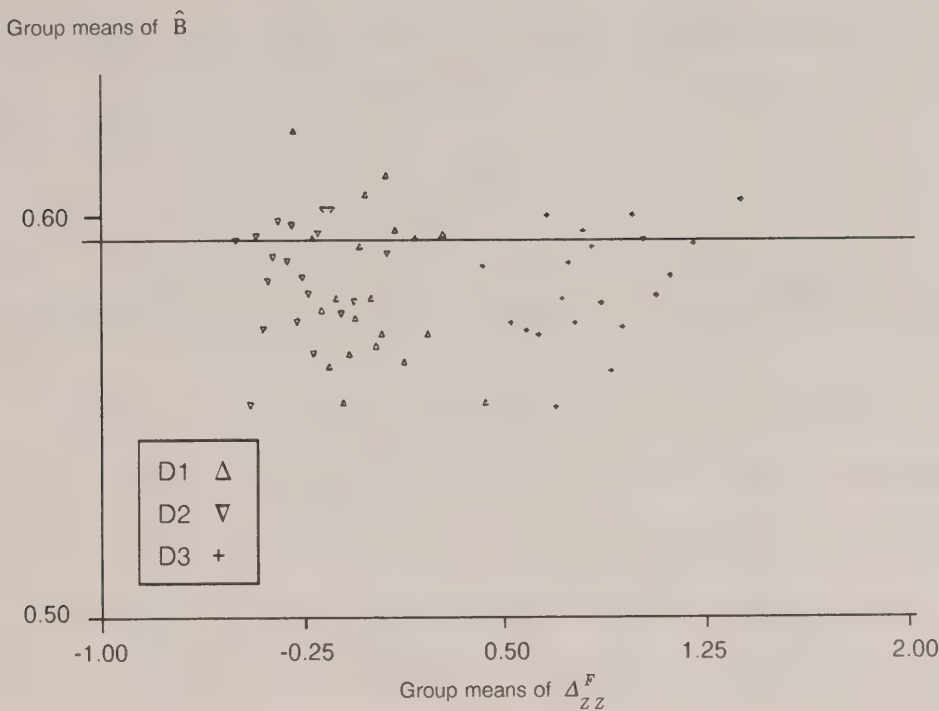


Figure 6.7 Scattergram of group means of $\hat{B}_{12,ml}$

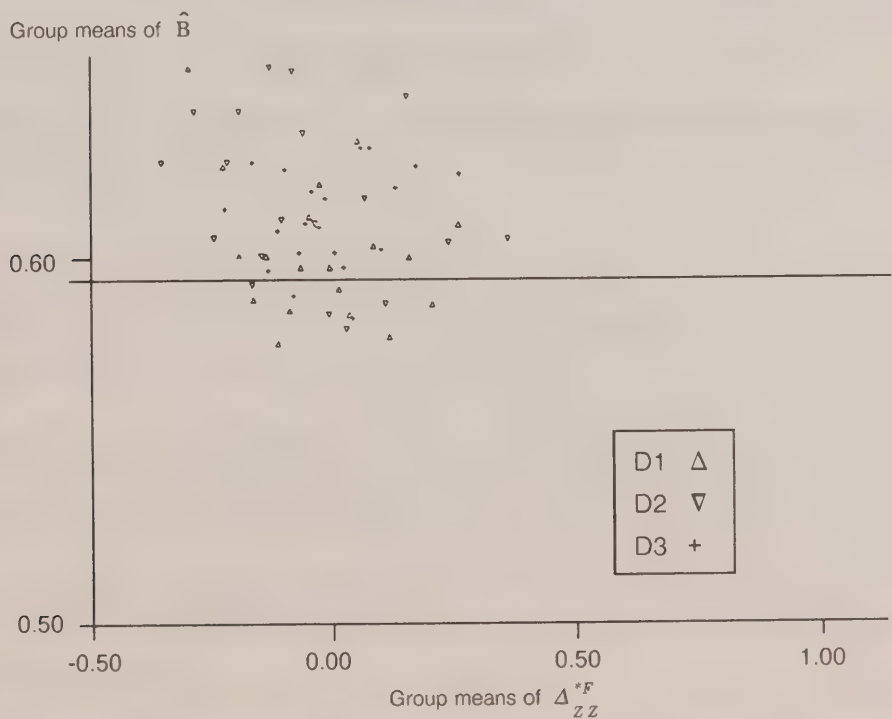


Figure 6.8 Scattergram of group means of $\hat{B}_{12,pwm}$

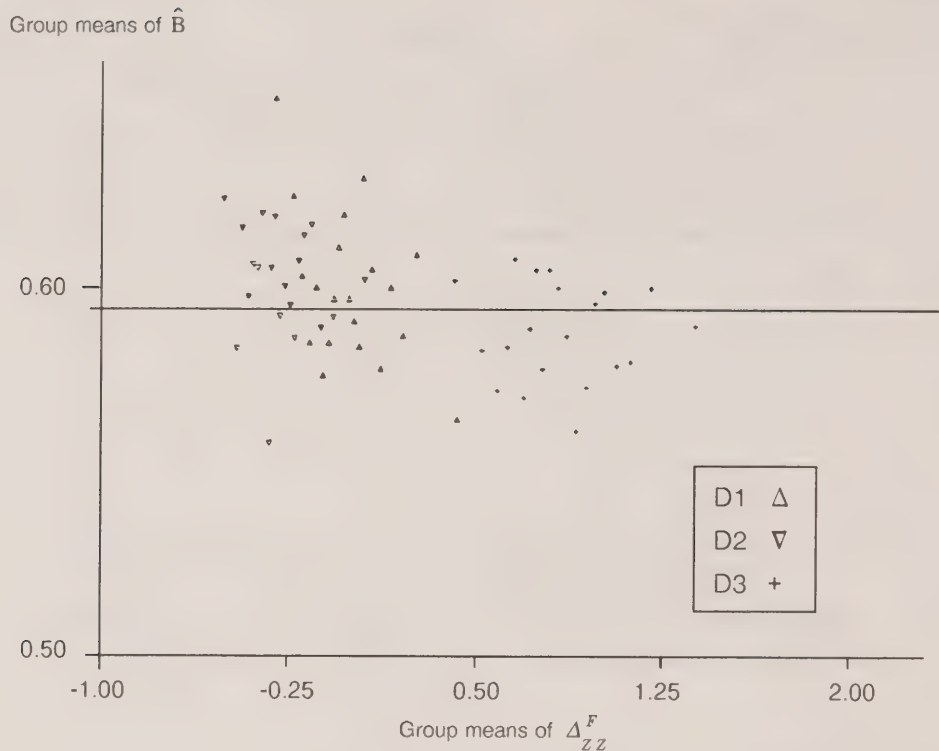


Figure 6.9 Scattergram of group means of $\hat{B}_{12,nw}$

ACKNOWLEDGEMENTS

The authors wish to thank an anonymous referee for many helpful comments which improved the presentation of the paper. E. Njenga was supported by a grant from the British Council.

REFERENCES

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BREWER, K.R.W. (1979). A class of robust designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W., and SÄRNDAL, C.-E. (1983). Six approaches to enumerative survey sampling. *Incomplete Data in Sample Surveys*, (Vol. 3). New York: Academic Press, 363-368.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GASSER, T., and MULLER, H.G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation*, (Eds. T. Gasser and M. Rosenblatt). New York: Springer-Verlag, 23-68.

GASSER, T., and ENGEL, J. (1990). The choice of weights in kernel regression estimation. *Biometrika*, 77, 377-381.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, B*, 17, 269-278.

- GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1211.
- GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.
- GODAMBE, V.P., and THOMPSON, M.E. (1977). Robust near optimal estimation in survey practice. *Bulletin of the International Statistical Institute*, 47, 129-146.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- HOLT, D., SMITH, T.M.F., and WINTERS, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Ser. A*, 143, 474-487.
- HOLMES, D. (1987). The effect of selection on the robustness of multivariate methods. Unpublished Doctoral thesis, University of Southampton, U.K.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTZ, S., and JOHNSON, N.L. (1988). *Encyclopedia of Statistical Sciences*, (Vol. 8). New York: John Wiley, 157.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability Application*, 9, 141-142.
- NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, B*, 42, 377-386.
- NJENGA, E.G. (1990). Robust estimation of the regression coefficients in complex surveys. Unpublished Ph.D. thesis, University of Southampton.
- PARZEN, E. (1962). On the estimation of the probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.
- PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions Royal Society of London, A*, 200, 1-66.
- PFEFFERMANN, D.J., and HOLMES, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, A*, 148, 268-278.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYALL, R.M., and HERSON, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- ROYALL, R.M. and HERSON, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association*, 68, 890-893.
- SÄRNDAL, C.-E. (1980). On π -inverse weighting versus best linear weighting in probability sampling. *Biometrika*, 67, 639-650.
- SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā, C*, 39, 1-9.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

- SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric curve fitting. *Journal of the Royal Statistical Society, B*, 47, 1-52.
- SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā, A*, 359-372.

Some Recent Work on Resampling Methods for Complex Surveys

J.N.K. RAO, C.F.J. WU and K. YUE¹

ABSTRACT

Resampling methods for inference with complex survey data include the jackknife, balanced repeated replication (BRR) and the bootstrap. We review some recent work on these methods for standard error and confidence interval estimation. Some empirical results for non-smooth statistics are also given.

KEY WORDS: Balanced repeated replication; Bootstrap; Jackknife; Stratified multistage designs; Variance estimation.

1. INTRODUCTION

Standard sampling theory is largely devoted to estimation of mean square error (MSE) of unbiased or approximately unbiased estimators \hat{Y} of a population total Y . An estimator of MSE, or a variance estimator, provides us with a measure of uncertainty in the estimator \hat{Y} . It is a common practice to assume that the estimator \hat{Y} is approximately normally distributed and then use a two-sided confidence interval $\hat{Y} \pm z_{\alpha/2}s(\hat{Y})$ or a one-sided confidence interval $(\hat{Y} - z_{\alpha}s(\hat{Y}), \infty)$ or $(-\infty, \hat{Y} + z_{\alpha}s(\hat{Y}))$, where $s(\hat{Y})$ is the standard error of \hat{Y} (*i.e.*, square root of estimated MSE) and z_{α} is the upper α -point of a $N(0, 1)$ variable. These intervals cover the true total Y with a probability of approximately $1 - \alpha$ in large samples, but the actual coverage probability could be significantly lower than $1 - \alpha$ in small samples or in highly clustered samples. For nonlinear statistics, such as ratios, regression or correlation coefficients, the well-known linearization (or Taylor expansion) method is often used (see Rao 1988 for detailed applications). Resampling methods, such as the jackknife, balanced repeated replication (BRR) and the bootstrap, are also being used, and in fact several agencies in the U.S.A and Canada have adopted the jackknife method of variance estimation for stratified multistage surveys. An advantage of the linearization method is that it is applicable to general sampling designs, but involves the derivation of a separate standard error formula, $s(\hat{\theta})$, for each nonlinear statistic, $\hat{\theta}$. On the other hand, resampling methods employ a single standard error formula for all statistics $\hat{\theta}$. However, the jackknife and the BRR methods are strictly applicable only to those stratified multistage designs in which clusters within strata are sampled with replacement or the first-stage sampling fraction is negligible. The bootstrap method of Rao and Wu (1987) works for more general designs, but it is computationally cumbersome and its properties for complex designs have not been fully investigated.

This paper provides an account of some recent work on resampling methods for complex surveys. Some empirical results on jackknife and bootstrap variance estimation for non-smooth statistics, such as the median, under stratified cluster sampling and stratified simple random sampling are also given.

¹ J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6.
C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1.
Kim Yue, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

2. STRATIFIED MULTISTAGE SAMPLING

Large-scale surveys often employ stratified multistage designs with large numbers of strata, L , and relatively few primary sampling units (clusters), $n_h (\geq 2)$, sampled within each stratum h . In fact, it is quite common to select $n_h = 2$ clusters within each stratum to permit maximum degree of stratification of clusters consistent with the provision of a valid variance estimator. We assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals Y_{hi} , $i = 1, \dots, n_h$; $h = 1, \dots, L$.

Let $w_{hik} (> 0)$ be the survey weight attached to the k -th sample element (ultimate unit) in the i -th sample cluster belonging to h -th stratum. Often, the basic weights w_{hik} are subjected to post-stratification adjustment to ensure consistency with known totals of post-stratification variables. For example, the Canadian Labour Force Survey uses a generalized regression estimator to ensure consistency. We shall, however, ignore this complication in the present paper. An estimator of the population total Y is of the form

$$\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}, \quad (2.1)$$

where s denotes the sample of elements and y_{hik} is the value of a characteristic of interest, y , associated with the sample element $(hik) \in s$. We assume complete response on all items.

It is a common practice to sample the clusters with probabilities proportional to sizes (pps) and without replacement to increase the efficiency of the estimators compared to pps sampling with replacement and to avoid the possibility of selecting the same cluster more than once in the sample. However, at the stage of variance estimation the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement and subsampling done independently each time a cluster is selected. This approximation leads to overestimation of variance of \hat{Y} , but the relative bias is likely to be small if the first stage sampling fraction is small in each stratum.

Writing \hat{Y} as

$$\hat{Y} = \sum_{h=1}^L \bar{r}_h, \quad (2.2)$$

with

$$r_{hi} = \sum_k (n_h w_{hik}) y_{hik}, \quad \bar{r}_h = \sum_i r_{hi} / n_h,$$

we note that the r_{hi} are independent and identically distributed (iid) random variables with the same mean, Y_h , and the same variance in each stratum h , under with replacement sampling of clusters. It therefore follows that an unbiased estimator of variance of \hat{Y} is given by

$$s^2(\hat{Y}) = \sum_h s_{rh}^2 / n_h, \quad (2.3)$$

with

$$(n_h - 1) s_{rh}^2 = \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2.$$

Under without-replacement sampling of clusters, $s^2(\hat{Y})$ will overestimate the true variance of \hat{Y} .

We are also often interested in estimating the population distribution function, $F(t)$, and the p -th quantile, $\theta = F^{-1}(p)$, $0 < p < 1$; in particular, the population median $\theta = F^{-1}(1/2)$. The survey estimator of $F(t)$ is given by

$$\hat{F}(t) = \sum_{(hik) \in s} \tilde{w}_{hik} a_{hik}, \quad (2.4)$$

where $\tilde{w}_{hik} = w_{hik} / \sum_s w_{hik}$ are the normalized weights ($\sum_s \tilde{w}_{hik} = 1$) and $a_{hik} = 1$ if $y_{hik} \leq t$, $a_{hik} = 0$ otherwise. The sample p -th quantile is obtained as

$$\hat{\theta} = \hat{F}^{-1}(p). \quad (2.5)$$

In practice, $\hat{\theta}$ is computed by first arranging the sampled values y_{hik} in an ascending order, say $\{y_{(hik)}\}$, and then cumulating the associated weights \tilde{w}_{hik} until p is first crossed. The first $y_{(hik)}$ encountered after crossing p is taken as the sample p -th quantile, $\hat{\theta}$. Woodruff (1952) obtained confidence intervals for a quantile, and Rao and Wu (1987) obtained a simple variance estimator using Woodruff's interval (see also Kovar, Rao and Wu 1988, Francisco and Fuller 1991). Shao (1991) considered general L -statistics, including the sample Lorenz curve and the Gini coefficient, which are examples of smooth L -statistics, and the sample quantiles which are examples of non-smooth L -statistics.

Many nonlinear parameters of interest, such as population means, ratios, regression and correlation coefficients, can be expressed as smooth functions, $\theta = g(Y)$, of a vector of totals, $Y = (Y_1, \dots, Y_q)'$, of suitably defined variates. An estimator of θ is given by $\hat{\theta} = g(\hat{Y})$. The linearization method may be used to estimate the variance of $g(\hat{Y})$, under any complex design (see Binder 1983 and Rao 1988).

3. RESAMPLING METHODS

Resampling methods, such as the jackknife and the bootstrap, are widely used in the iid case. Suitable modification/extensions of these methods have also been developed to handle survey data involving stratification and clustering. We now give a brief account of some recent work on three such methods: jackknife, balanced repeated replication and bootstrap, in the context of stratified multistage sampling.

3.1 Jackknife

For simplicity, assume $\hat{\theta} = g(\hat{Y})$, a smooth function of the estimated total \hat{Y} . Let $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$ be the estimator of θ obtained from the sample after omitting the data from the j -th sampled cluster in g -th stratum ($j = 1, \dots, n_g$; $g = 1, \dots, L$), where

$$\hat{Y}_{(gj)} = \sum_{\substack{(hik) \in s \\ h \neq g}} w_{hik} y_{hik} + \sum_{\substack{(gik) \in s \\ i \neq j}} \left\{ \frac{n_g}{n_g - 1} w_{gik} \right\} y_{gik}. \quad (3.1)$$

Note that $\hat{Y}_{(gj)}$ is obtained by changing the weight of (gik) -th element to $n_g w_{gik} / (n_g - 1)$, $i \neq j$, but retaining the original weights, w_{hik} , for $h \neq g$. A customary delete-1 cluster jackknife variance estimator of $\hat{\theta}$ is given by

$$s_J^2(\hat{\theta}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2. \quad (3.2)$$

Two variations of $s_J^2(\hat{\theta})$ are obtained by changing $\hat{\theta}$ in (3.2) to $\hat{\theta}_{(g\cdot)} = \sum_j \hat{\theta}_{(gj)}/n_g$ and $\hat{\theta}_{(\cdot\cdot)} = \sum_g \sum \hat{\theta}_{(gj)}/n$, where $n = \sum_g n_g$. In the linear case, $\hat{\theta} = \hat{Y}$, all the jackknife variance estimators reduce to the “correct” variance estimator, $s^2(\hat{Y})$, given by (2.3). Rao and Wu (1987) made a second order analysis of the resampling variance estimators when $\hat{\theta}$ is expressed as a smooth function of totals, \hat{Y} . Their main results on the jackknife are: (1) Different jackknife variance estimators are asymptotically equal to higher order terms, as the number of strata, L , increases. (2) In the important case of $n_h = 2$ for all h , the linearization variance estimator, $s_L^2(\hat{\theta})$, and any jackknife variance estimator are asymptotically equal to higher order terms, indicating that the choice between the two methods should depend more on operational considerations than on statistical criteria.

A drawback of the customary delete-1 jackknife method in the case of independent and identically distributed (i.i.d.) observations is that, unlike the bootstrap, it fails to provide a consistent variance estimator for non-smooth statistics, such as the median. Shao and Wu (1989), however, have shown that this deficiency of the delete-1 jackknife can be rectified by using a more general jackknife, called the delete- d jackknife, with the number of observations deleted, d , depending on a smoothness measure of the statistic. In particular, for the sample quantiles, the delete- d jackknife with d satisfying $n^{1/2}/d \rightarrow 0$ and $n - d \rightarrow \infty$ as $n \rightarrow \infty$ leads to consistent variance estimators in the case of i.i.d. observations. This result suggests that a similar effect might hold in the case of delete-1 cluster jackknife for stratified multistage sampling since all the sampled elements in a sampled cluster (gj) are deleted in computing $s_J^2(\hat{\theta})$ given by (3.2). At present we are studying this problem theoretically, but we performed a limited simulation study which suggests that the delete-1 cluster jackknife variance estimator $s_J^2(\hat{\theta})$ might perform quite well. We now report the results of the simulation study for the median, $\hat{\theta} = \hat{F}^{-1}(1/2)$.

For the simulation study, we generated stratified cluster samples $\{y_{hik}, k = 1, \dots, M; i = 1, \dots, n_h; h = 1, \dots, L\}$ employing the nested error model $y_{hik} = \mu_h + a_{hi} + e_{hik}$ with $a_{hi} \stackrel{iid}{\sim} N(0, \sigma_{ah}^2)$ and $e_{hik} \stackrel{iid}{\sim} N(0, \sigma_{eh}^2)$, where the cluster size, M is assumed to be equal for all clusters (hi), and the intra-cluster correlations, $\sigma_{ah}^2/(\sigma_{ah}^2 + \sigma_{eh}^2) = \rho_h$, are assumed to be equal for all strata h (i.e., $\rho_h = \rho$). The normalized survey weights are given by \tilde{w}_{hik} with $w_{hik} = W_h/(n_h M)$ and W_h denotes the relative size of stratum h . The number of strata $L (= 32)$, strata means, μ_h , variances $\sigma_h^2 = \sigma_{ah}^2 + \sigma_{eh}^2$ and sizes W_h were chosen to correspond to real populations encountered in the US National Assessment of Educational Progress Study (Hansen and Tepping 1985). We generated 1,000 independent stratified cluster samples with $n_h = 2$ for each selected combination (ρ, M) and then computed the bias and relative bias of the jackknife variance estimator, $s_J^2(\hat{\theta})$, for the median: $\text{Bias}[s_J^2(\hat{\theta})] = \sum_t s_{Jt}^2(\hat{\theta})/1,000 - \text{MSE}(\hat{\theta})$, where $s_{Jt}^2(\hat{\theta})$ is the value of $s_J^2(\hat{\theta})$ for the t -th simulated sample ($t = 1, \dots, 1,000$) and $\text{Rel. Bias}[s_J^2(\hat{\theta})] = \text{Bias}[s_J^2(\hat{\theta})]/\text{MSE}(\hat{\theta})$. We calculated $\text{MSE}(\hat{\theta})$ from an independent set of 10,000 stratified cluster samples for each (ρ, M) : $\text{MSE}(\hat{\theta}) = \sum_t (\hat{\theta}_t - \hat{\theta})^2/10,000$, where $\hat{\theta}_t$ is the value of $\hat{\theta}$ for the t -th simulated sample, $\hat{\theta} = \sum \hat{\theta}_t/10,000$ and $t = 1, \dots, 10,000$.

Table 1 reports the simulated values of bias and relative bias (in brackets) of the jackknife variance estimator for selected combinations of ρ and M . First, we note that for the special case of stratified simple random sampling ($\rho = 0, M = 1$), the relative bias is very large (116%) thus confirming the inconsistency of $s_J^2(\hat{\theta})$ in this case. Second, we observe that both the bias and relative bias decrease as M increases for a given ρ . Moreover, for a given cluster

Table 1
Bias and % Relative Bias (in Brackets) of Jackknife Variance Estimator for
the Median Under Stratified Cluster Sampling ($n_h = 2, L = 32$)
and Selected Values of Equal Intra-Cluster Correlation, ρ ,
and Equal Cluster Size, M

ρ	M				
	1	10	20	30	50
0	7.5(116)	.28(41)	.09(29)	.04(15)	.01(15)
0.05	–	.22(27)	.09(18)	.05(12)	.03 (8)
0.10	–	.28(28)	.10(14)	.06 (9)	.02 (3)
0.20	–	.31(22)	.11(10)	.08 (8)	.03 (3)
0.30	–	.32(18)	.11 (7)	.07 (5)	.01 (1)
0.50	–	.44(17)	.15 (6)	.11 (5)	.04 (2)

size M , the bias generally increases with ρ , but the relative bias in fact decreases because $\text{MSE}(\hat{\theta})$ is increasing faster than the bias as ρ increases. It is indeed gratifying that the relative bias is no more than 10% for $M \geq 30$ and $\rho \geq 0.10$ or $M \geq 20$ and $\rho \geq 0.20$.

3.2 Balanced Repeated Replication (BRR)

Balanced repeated replication (BRR) was proposed by McCarthy (1969) for the important special case of $n_h = 2$ clusters per stratum. A set of R balanced half-samples (replications) is formed by deleting one cluster from the sample in each stratum. This set may be defined by a $R \times L$ design matrix (δ_{rh}^r) , $1 \leq r \leq R, 1 \leq h \leq L$ with $\delta_{rh}^r = +1$ or -1 according as whether the first or second sample cluster in the h -th stratum is in the r -th half-sample, and $\sum_r \delta_{rh}^r \delta_{h'}^r = 0$ for all $h \neq h'$, i.e. the columns of the matrix are orthogonal. A minimal set of R balanced half-samples may be constructed from Hadamard matrices ($L + 1 \leq R \leq L + 4$) by choosing any L columns, excluding the column of $+1$'s.

Let $\hat{\theta}^{(r)}$ be the estimator of θ obtained from the r -th half-sample. Note that $\hat{\theta}^{(r)}$ is obtained from $\hat{\theta}$ by changing the weight of (hik) -th element to $2w_{hik}$ or 0 according as the (hi) -th cluster is selected or not selected in the half-sample. A BRR variance estimator of $\hat{\theta}$ is given by

$$s_{\text{BRR}}^2(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.3}$$

Several variations of $s_{\text{BRR}}^2(\hat{\theta})$ are also available; for example, $\hat{\theta}$ may be changed to $\hat{\theta}(\cdot) = \sum_r \hat{\theta}^{(r)}/R$. In the linear case, $\hat{\theta} = \hat{Y}$, all the BRR variance estimators reduce to the “correct” variance estimator, $s^2(\hat{Y})$, as in the case of the jackknife.

Krewski and Rao (1981) established the consistency of $s_J^2(\hat{\theta})$ and $s_{\text{BRR}}^2(\hat{\theta})$ for smooth statistics $\hat{\theta} = g(\hat{Y})$, as L increases. Rao and Wu (1985) made a second order analysis and showed that $s_{\text{BRR}}^2(\hat{\theta})$ and $s_L^2(\hat{\theta})$ are not asymptotically equivalent to second order terms, unlike $s_J^2(\hat{\theta})$ and $s_L^2(\hat{\theta})$. Shao and Wu (1992) established the consistency of $s_{\text{BRR}}^2(\hat{\theta})$ for the quantiles, $\hat{\theta} = \hat{F}^{-1}(p)$.

The BRR method has been extended to the case of $n_h = p > 2$ clusters per stratum for p prime or power of prime (Gurney and Jewett 1975), but the number of replications, R , needed is much larger than in the case of $n_h = 2$. In many survey designs n_h 's are not equal. To accommodate the general case of unequal n_h , Gupta and Nigam (1987) and Wu (1991)

advocated the use of mixed-level orthogonal arrays of strength two for drawing balanced replicates, where n_h is the number of symbols in the h -th column of the array. Orthogonality of the array guarantees that the replicates drawn are balanced. Unlike the case of equal n_h , the adjustment of survey weights is more complicated. A correct method was given by Wu (1991). From his formula (6), two separate adjustments should be applied to the sampled and unsampled units in each replicate. Simple algebra on Wu's equation (6) shows that w_{hik} is changed to $w'_{hik} = [1 + (n_h - 1)^{1/2}] w_{hik}$ or $w''_{hik} = [1 - (n_h - 1)^{1/2}] w_{hik}$ according as the (hik) -th element is selected or not selected in the replicate. (Note that $w'_{hik} = 2$ and $w''_{hik} = 0$ for $n_h = 2$). The remaining calculation of $\hat{\theta}^{(r)}$ and $s^2_{BRR}(\hat{\theta})$ are the same as in (3.3). Furthermore, these modified survey weights can be applied to $\hat{\theta} = \hat{F}^{-1}(p)$ and more general $\hat{\theta} = T(\hat{F})$, where T is a functional of \hat{F} . All we need to do is to change w_{hik} in (2.4) to w'_{hik} or w''_{hik} according as the (hik) -th element is selected or not selected in the r -th replicate to get $\hat{F}^{(r)}$ of F for the r -th replicate, and $\hat{\theta}^{(r)} = T(\hat{F}^{(r)})$. The calculation of the BRR variance estimator is the same as in (3.3).

There are two problems with the use of mixed orthogonal arrays. First, the array size can be large for general n_h . Second, orthogonal arrays do not exist for any combination of n_h 's. A practical solution is to group the n_h sample psu's in stratum h into two to four groups of psu's and then apply the method to the groups by treating the groups as units in the BRR method. This extension is called the grouped BRR method. As shown by Wu (1991), its efficiency loss can be relatively small, compared to the full BRR, if the groupings are done judiciously. For example, more groups are needed if n_h is large and the units within the stratum are more heterogeneous. For $n_h = 2, 3$ or 4 , many mixed orthogonal arrays have been constructed (see, for example, Dey 1985 and Wang and Wu 1991). If n_h can only take 2 or 4, saturated orthogonal arrays for any combination can be easily constructed as in Wu (1989). That is, the number of replications can be as small as possible. It is therefore possible to compile a large collection of mixed orthogonal arrays for practical use if n_h is restricted to 2, 3 or 4.

The BRR method and extensions considered thus far only take one unit (psu) per stratum for each replicate. If n_h is large, say more than 3, Sitter (1992) proposed the use of orthogonal multi-arrays to allow the number of resampled units per stratum to be greater than one. It may require fewer replicates and it can cover cases where orthogonal arrays of strength two are not available; for example, $n_h = 6$.

3.3 Bootstrap

The bootstrap method for the iid case has been extensively studied (Efron 1982). Rao and Wu (1987) provided an extension to stratified multistage designs, but covering only smooth statistics $\hat{\theta} = g(\hat{Y})$. They required that, in order to have valid variance estimation in the case of small n_h , some scale adjustment, similar to those in Section 3.2, is necessary. What they did not realize is that the scale adjustment should be made on the survey weights w_{hik} rather than the y_{hik} values directly, which is what they proposed. As a result, their method cannot be extended to cover the quantile $\theta = F^{-1}(p)$. We now present a general method that covers smooth as well as non-smooth statistics for arbitrary sizes, n_h . It works as follows: (i) Draw a simple random sample of m_h clusters with replacement from the n_h sample clusters, independently for each h . Let m^*_{hi} be the number of times (hi) -th sample cluster is selected ($\sum_i m^*_{hi} = m_h$). Define the bootstrap weights

$$w^*_{hik} = \left[\{1 - (m_h / (n_h - 1))^{1/2}\} + (m_h / (n_h - 1))^{1/2} (n_h / m_h) m^*_{hi} \right] w_{hik}. \quad (3.4)$$

If the (hi) -th cluster is not selected in the bootstrap sample, $m_{hi}^* = 0$ and the second term of (3.4) vanishes. If m_h is chosen to be less than or equal to $n_h - 1$, then the bootstrap weights w_{hik}^* are all positive if $w_{hik} > 0$ for all $(hik) \in s$. Calculate θ^* , the bootstrap estimator of θ , using the weights w_{hik}^* in the formula for $\hat{\theta}$. The bootstrap median, for example, is calculated as before using the normalized bootstrap weights $\tilde{w}_{hik}^* = w_{hik}^* / \sum_s w_{hik}^*$, provided all $w_{hik}^* > 0$. (ii) Independently replicate step (i) a large number, B , of times and calculate the corresponding estimates $\theta_{(1)}^*, \dots, \theta_{(B)}^*$.

The bootstrap variance estimator $s_{\text{BOOT}}^2(\hat{\theta}) = E_*(\theta^* - E_*\theta^*)^2$, is approximated by

$$\hat{s}_{\text{BOOT}}^2(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B [\theta_{(b)}^* - \hat{\theta}]^2. \tag{3.5}$$

A variation of (3.5) is obtained by changing $\hat{\theta}$ to $\theta_{(\cdot)}^* = \sum_b \theta_{(b)}^* / B$. In the linear case, $s_{\text{BOOT}}^2(\hat{\theta})$ reduces to the "correct" variance estimator $s^2(\hat{Y})$.

Rao and Wu (1987) obtained bootstrap- t confidence intervals for smooth functions, $\theta = g(Y)$, by approximating the distribution of $t = (\hat{\theta} - \theta) / s_J(\hat{\theta})$ by its bootstrap counterpart $t^* = (\theta^* - \hat{\theta}) / s_J(\theta^*)$, where $s_J^2(\theta^*)$ is obtained from (3.2) with w_{hik} changed to w_{hik}^* . A two-sided $(1 - \alpha)$ -level confidence interval for θ is then given by $\{\hat{\theta} - t_L^* s_J(\hat{\theta}), \hat{\theta} - t_U^* s_J(\hat{\theta})\}$, where t_L^* and t_U^* are the lower and upper $\alpha/2$ -points of t^* obtained from the bootstrap histogram of $t_{(1)}^*, \dots, t_{(B)}^*$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Empirical work by Kovar, Rao and Wu (1988) for smooth functions indicates that the bootstrap- t interval with $m_h = n_h - 1$ tracks the error rates in both the lower and upper tails better than the jackknife interval $\{\hat{\theta} - z_{\alpha/2} s_J(\hat{\theta}), \hat{\theta} + z_{\alpha/2} s_J(\hat{\theta})\}$, but the total error rate is not distinguishable from the latter, i.e., for two-sided intervals, they exhibit similar performance in terms of actual coverage probability. If a variance stabilizing transformation can be found, such as the \tanh^{-1} transformation on the estimated correlation coefficient, then the problem of uneven error rates in the two tails for the jackknife interval seems to be corrected. This suggests that the jackknife interval, or any other normal-theory interval, based on such transformations can be useful when the transformations are known, while the bootstrap provides an alternative when such transformations do not exist or are unknown.

We now present the results of a limited simulation study on the performance of the proposed bootstrap method in the case of the median. Employing the Hansen-Tepping basic population 1 with $L = 32$ strata (see Kovar *et al.* 1988, Sections 3 and 6 for details), we generated 500 independent stratified simple random samples with $n_h = 5$ and then computed the relative bias and coefficient of variation (relative stability) of the Woodruff-based variance estimator with $\alpha = 0.1$ (see Kovar *et al.* 1988, eq. (2.8)), the BRR variance estimator (3.3) and the bootstrap variance estimator (3.5) and its variation obtained by changing $\hat{\theta}$ to $\theta_{(\cdot)}^*$. We used $m_h = n_h - 1$ and $n_h - 3$ and $B = 500$ bootstrap replicates for each sample, while the BRR replicates were obtained from an orthogonal array with 250 runs. The true MSE of $\hat{\theta}$ was approximated by selecting 10,000 independent stratified random samples. We also calculated the error rates in each tail (nominal rate of 5% in each tail) and standardized lengths of the normality-based confidence interval using the BRR variance estimator, the Woodruff interval and the bootstrap interval obtained from the percentile method using the bootstrap histogram of $\theta_{(1)}^*, \dots, \theta_{(B)}^*$ for each sample.

Table 2 reports the simulated values of the relative bias, coefficient of variation, lower (L) and upper (U) error rates, and standardized lengths. First, we note that the bootstrap variance estimator (3.5) has a larger relative bias and a slightly larger coefficient of variation (CV) than

Table 2
% Relative Bias and % CV of Variance Estimator and Error Rates
and Standardized Lengths of Confidence Intervals
(Nominal Level of 5% in Each Tail) for the Median Under Stratified
Simple Random Sampling $L = 32, n_h = 5$

Method	% Rel. Bias	% CV	Error Rate		St. Length
			L	U	
Woodruff	4.2	47	4.2	5.6	0.997
BRR	3.1	31	5.0	5.0	1.004
Bootstrap*:					
$m_h = 4$	12.6 (7.5)	52 (48)	5.0	5.2	0.987
$m_h = 2$	13.0 (7.8)	54 (49)	5.0	4.8	0.988

* Results for the variation of the bootstrap variance estimator are given in the brackets.

its variation obtained by changing $\hat{\theta}$ to $\theta^*_{(.)}$: Relative bias of 12.6% vs. 7.5% and CV of 52% vs. 48% for $m_h = n_h - 1 = 4$. On the other hand, the BRR variance estimator has the smallest relative bias (3.1%) and the smallest CV (31%), while the Woodruff-based variance estimator has a smaller relative bias (4.2%) and a comparable CV (47%). Secondly, the lower and upper error rates are close to the nominal level (5%) for the bootstrap and the BRR intervals, while the error rates are slightly uneven for the Woodruff interval ($L = 4.2\%$ and $U = 5.6\%$). Finally, we note that the standardized lengths are roughly equal for all the methods. Overall, the bootstrap variance estimator and the bootstrap intervals based on the percentile method did not exhibit better performance relative to either the BRR variance estimator and the associated normality-based interval or the Woodruff-based variance estimator and the Woodruff interval.

ACKNOWLEDGEMENT

J.N.K. Rao’s work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

DEY, A. (1985). *Orthogonal Fractional Factorial Designs*. New Delhi: Wiley Eastern.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

GURNEY, M., and JEWETT, R.S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, 70, 819-821.

- HANSEN, M., and TEPPING, B.J. (1985). Estimation for variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.
- KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25-45.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.
- RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics, Vol. 6*, (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: Elsevier Science, 427-447.
- RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- RAO, J.N.K., and WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- RAO, J.N.K., and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data. *Bulletin of the International Statistical Institute*.
- SHAO, J. (1991). *L*-statistics in complex survey problems. Technical Report, University of Ottawa, Ottawa.
- SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.
- SHAO, J., and WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics*, 20 (to appear).
- SITTER, R.R. (1992). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, (to appear).
- WANG, J.C., and WU, C.F.J. (1991). An approach to the construction of asymmetrical orthogonal arrays. *Journal of the American Statistical Association*, 86, 450-456.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other positional measures. *Journal of the American Statistical Association*, 47, 635-646.
- WU, C.F.J. (1989). Construction of $2^m 4^n$ designs via a grouping scheme. *Annals of Statistics*, 17, 1880-1885.
- WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

An Estimating Function Approach to Finite Population Estimation

HAROLD J. MANTEL¹

ABSTRACT

Godambe and Thompson (1986) define and develop simultaneous optimal estimation of superpopulation and finite population parameters based on a superpopulation model and a survey sampling design. Their theory defines the finite population parameter, θ_N , as the solution of the optimal estimating equation for the superpopulation parameter θ ; however, some other finite population parameter, ϕ , may be of interest. We propose to extend the superpopulation model in such a way that the parameter of interest, ϕ , is a known function of θ_N , say $\phi = f(\theta_N)$. Then ϕ is optimally estimated by $f(\theta_s)$, where θ_s is the optimal estimator of θ_N , as given by Godambe and Thompson (1986), based on the sample s and the sampling design.

KEY WORDS: Estimating functions; Generalized linear estimator; Finite population parameter.

1. ESTIMATION OF A MEAN

The problem discussed in this paper is the estimation of a finite population parameter such as the mean based on a sample survey. There is also a hypothesized superpopulation regression model relating the variable of interest to some known covariables. The objective is an estimation procedure which has good properties with respect to both the sampling design and the hypothesized model. The approach here is based on the work of Godambe and Thompson (1986).

We suppose that we have a finite population of labeled individuals $P = \{i: i = 1, \dots, N\}$. With each individual i is associated an unknown variable y_i and a vector of covariables, \mathbf{x}_i . The vector \mathbf{x}_i may be known for all $i \in P$ or only for i in the sample and the population mean $\bar{\mathbf{x}}_N$ would be known. Letting E_m denote expectation with respect to the superpopulation model, the model assumptions are:

- (i) y_i and y_j are independent for $i \neq j$
- (ii) $E_m(y_i) = \mathbf{x}_i^T \beta$ for some unknown real vector β
- (iii) $E_m(y_i - \mathbf{x}_i^T \beta)^2 = \sigma^2 v_i$, $i = 1, \dots, N$, for known v_i and some unknown σ^2 .

Following Godambe and Thompson (1986) we define a finite population parameter $\hat{\beta}_N$ as the solution of the linearly optimal estimating equation

$$\mathbf{g}^* = \sum_{i=1}^N (y_i - \mathbf{x}_i^T \beta) \mathbf{x}_i / v_i = 0, \quad (1)$$

that is,

$$\hat{\beta}_N = (\mathbf{X}_N^T \mathbf{V}_N^{-1} \mathbf{X}_N)^{-1} \mathbf{X}_N^T \mathbf{V}_N^{-1} \mathbf{y}_N, \quad (2)$$

¹ H.J. Mantel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

where $y_N^T = (y_1, \dots, y_N)$, V_N is a diagonal matrix with entries v_1, \dots, v_N , and X_N is a matrix with N rows, the i th row being x_i^T .

Now $\hat{\beta}_N$ is unknown. Godambe and Thompson (1986) defined and developed simultaneous optimal estimation of β and $\hat{\beta}_N$ based on the model and the sampling design. We will denote the data from a sample survey by $\chi_s = \{ (i, y_i), i \in s \}$.

For simultaneous estimation of β and $\hat{\beta}_N$ we consider estimating functions $h(\chi_s, \beta)$ such that $E_p(h) = g^*$ in (1), where E_p denotes expectation with respect to the sampling design. A function h^* in this class is called optimal if for all other h in the class $E_m E_p \{ h h^T \} - E_m E_p \{ h^* h^{*T} \}$ is non-negative definite. Theorem 1 of Godambe and Thompson (1986) shows that the optimal function h^* is given by

$$h^*(\chi_s, \beta) = \sum_{i \in s} (y_i - x_i^T \beta) x_i / \pi_i v_i, \tag{3}$$

where π_i is the probability under the sampling design that individual i is included in the sample s . We will denote the root of this function by $\hat{\beta}_s$, that is,

$$\hat{\beta}_s = (X_s^T \Pi_s^{-1} V_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} V_s^{-1} y_s, \tag{4}$$

where y_s is the vector of y_i s for $i \in s$, Π_s and V_s are diagonal matrices with entries π_i and v_i respectively, $i \in s$, and X_s is the matrix with rows x_i^T , $i \in s$.

So far we have discussed only estimation of β or $\hat{\beta}_N$. Our problem was to estimate \bar{y}_N , the population mean of the y_i s. One possibility is to use a generalized regression estimator,

$$\bar{y}_{\text{GREG}} = \bar{x}_N^T \hat{\beta}_s + \mathbf{1}_s^T \Pi_s^{-1} (y_s - X_s \hat{\beta}_s) / N, \tag{5}$$

where $\mathbf{1}_s$ is a vector of 1's whose length is the size of the sample s . This estimator is discussed, for example, by Särndal, Swensson and Wretman (1992). The first part of the estimator gives good model properties while the second part gives good design properties. However, the model and design justifications of \bar{y}_{GREG} in (5) do not depend on the particular form of $\hat{\beta}_s$, and there is no immediately apparent reason why $\hat{\beta}_s$ in (5) could not be replaced by a purely model based estimator of β . The design optimality of $\hat{\beta}_s$ is apparently irrelevant.

The estimator we will propose here more closely integrates the hypothesized model with the finite population parameter \bar{y}_N . Since $\hat{\beta}_N$ in (2) is optimally estimated by $\hat{\beta}_s$ in (4), functions of $\hat{\beta}_N$ are optimally estimated by the same function of $\hat{\beta}_s$. If $\bar{y}_N = u^T \hat{\beta}_N$ for some vector u then we would estimate \bar{y}_N by $u^T \hat{\beta}_s$. Such a u exists if and only if $V_N \mathbf{1}_N$ is in the column space of X_N , in which case, with $V_N \mathbf{1}_N = X_N a$, we may take $u = X_N^T V_N^{-1} X_N a / N = \bar{x}_N$. The idea then is that if $V_N \mathbf{1}_N$ is not in the column space of X_N , we will add it. In doing so we lose something of model efficiency, though the augmented model remains valid in light of the original model. We relax model efficiency to gain some sort of finite population relevance. As an interesting special case we note that when the model variances do not depend on i our approach leads to including an arbitrary constant term in the regression model.

The approach taken here seems quite similar to that of Little (1983) who suggests model based estimation restricting attention to models that yield asymptotically design consistent estimators. Alternatively, Isaki and Fuller (1982) suggest restricting to designs for which the model based estimator is asymptotically design consistent.

2. COMPARISON TO THE GENERALIZED REGRESSION ESTIMATOR

Let W_N be the design matrix for the augmented model, that is

$$W_N = (V_N \mathbf{1}_N, X_N). \quad (6)$$

For the discussion of this section we assume that $V_N \mathbf{1}_N$ is not in the column space of X_N . Similarly, let W_s be the augmented form of X_s , and γ , $\hat{\gamma}_N$, and $\hat{\gamma}_s$ be the augmented forms of β , $\hat{\beta}_N$, and $\hat{\beta}_s$ respectively.

For convenience, we will refer to our estimator of the population mean as the augmented regression estimator,

$$\bar{y}_{\text{AREG}} = \bar{w}_N^T \hat{\gamma}_s. \quad (7)$$

We first show that \bar{y}_{AREG} is also a type of generalized difference estimator. From (6), if \mathbf{u} is a vector of appropriate length with the first entry equal to one and the rest zeros then $W_N \mathbf{u} = V_N \mathbf{1}_N$ and $W_s \mathbf{u} = V_s \mathbf{1}_s$. Then

$$\mathbf{1}_s^T \Pi_s^{-1} W_s \hat{\gamma}_s = \mathbf{u}^T W_s^T V_s^{-1} \Pi_s^{-1} W_s \hat{\gamma}_s = \mathbf{u}^T W_s^T V_s^{-1} \Pi_s^{-1} \mathbf{y}_s = \mathbf{1}_s^T \Pi_s^{-1} \mathbf{y}_s$$

and it follows that the second part of the generalized regression estimator in (5) with $\hat{\beta}_s$ replaced by $\hat{\gamma}_s$ is equal to 0.

Secondly, let us compare \bar{y}_{AREG} in (7) to \bar{y}_{GREG} in (5). A few tedious calculations give us that

$$\bar{y}_{\text{AREG}} = \bar{x}_N \hat{\beta}_s + (c_1/c_2) \mathbf{1}_s^T \Pi_s^{-1} (\mathbf{y}_s - X_s \hat{\beta}_s) / N,$$

where

$$c_1 = \mathbf{1}_N^T (V_N \mathbf{1}_N - X_N (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s)$$

and

$$c_2 = \mathbf{1}_s^T \Pi_s^{-1} (V_s \mathbf{1}_s - X_s (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s).$$

Written in this way \bar{y}_{AREG} appears very similar to \bar{y}_{GREG} except for an adjusted weight for the second part. It does not seem possible to give an heuristic explanation of the weight (c_1/c_2) . However, we note that c_1 is just the population sum of the residuals from a weighted regression of the v_i 's onto the x_i 's based on the sample s , and c_2 looks something like a Horvitz-Thompson estimator of c_1 , except that the residuals also depend on the sample s . For large samples from large populations we would expect (c_1/c_2) to be close to 1.

In comparing \bar{y}_{AREG} with \bar{y}_{GREG} we may say that \bar{y}_{AREG} is more design based and \bar{y}_{GREG} is more model based. Of course, \bar{y}_{GREG} is design consistent, but \bar{y}_{AREG} has also a finite sample design justification in that $\hat{\gamma}_s$ is the solution of an estimating equation which is design unbiased for the parameter defining equation of β_N . Parameter defining equations are discussed by Godambe and Thompson (1984, 1986).

3. VARIANCE ESTIMATION AND CONFIDENCE INTERVALS

A method of confidence interval construction which would be consistent with the general philosophy of estimating functions would be to construct an asymptotically multivariate normal pivotal based on \mathbf{h}^* and an estimator of its variance. Approximate confidence regions for $\hat{\gamma}_N$ would then correspond to probability regions of the estimated multivariate normal distribution of this approximate pivotal. However, we are not interested in $\hat{\gamma}_N$ but in a non-injective function of $\hat{\gamma}_N$. We will adopt the more straight-forward approach of estimating the variance of \bar{y}_{AREG} directly.

Särndal, Swensson, and Wretman (1989) have investigated variance estimation for \bar{y}_{GREG} in (5) for the case that the second part is zero. As we have seen in section 2, our estimator \bar{y}_{AREG} is precisely of that type. Their variance estimator may be written as

$$\hat{V}_g = \sum_{i \in s} \sum_{j \in s} \tilde{\Delta}_{ij} g_{is} \tilde{e}_{is} g_{js} \tilde{e}_{js}, \quad (8)$$

where $\tilde{\Delta}_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij}$, π_{ij} is the design probability that both individuals i and j are included in the sample s , g_{is} is the i th element of the row vector $\bar{\mathbf{w}}_N^T (W_s^T V_s^{-1} \Pi_s^{-1} W_s)^{-1} W_s^T V_s^{-1}$, and $\tilde{e}_{is} = (y_i - x_i^T \hat{\gamma}_s) / \pi_i$. See Särndal, Swensson and Wretman (1989) for a detailed discussion of the model and design properties of \hat{V}_g in (8). Note that \bar{y}_{AREG} in (7) may be written as $\bar{y}_{\text{AREG}} = \sum_{i \in s} g_{is} y_i / \pi_i$ and

$$\bar{y}_{\text{AREG}} - \bar{y}_N = \sum_{i \in s} g_{is} \tilde{e}_{iN} = \bar{\mathbf{w}}_N^T (\hat{\gamma}_s - \hat{\gamma}_N),$$

where $\tilde{e}_{iN} = (y_i - \mathbf{w}_i^T \hat{\gamma}_N) / \pi_i$. Now, with $V_N \mathbf{1}_N = W_N \mathbf{a}$, we have $\bar{\mathbf{w}}_N^T = \mathbf{1}_N^T V_N V_N^{-1} W_N / N = \mathbf{a}^T W_N^T V_N^{-1} W_N / N$, so that for large samples g_{is} will be near $1/N$ for $i \in s$. The design variance of \bar{y}_{AREG} is then approximately equal to

$$\sum_{i \in P} \sum_{j \in P} \Delta_{ij} \tilde{e}_{iN} \tilde{e}_{jN} / N^2,$$

where $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)$, and this may be estimated by

$$\hat{V}_1 = \sum_{i \in s} \sum_{j \in s} \tilde{\Delta}_{ij} \tilde{e}_{is} \tilde{e}_{js} / N^2. \quad (9)$$

\hat{V}_1 in (9) was considered in early work on the general regression estimator, for example, Särndal (1981, 1982). Now \hat{V}_g in (8) may be thought of as a version of \hat{V}_1 in (9) adjusted for the realized values of g_{is} , $i \in s$. Särndal, Swensson and Wretman (1989) show that \hat{V}_g in (8), as well as being design consistent for the design variance of \bar{y}_{AREG} , is often model unbiased or nearly model unbiased for the model mean squared error of \bar{y}_{AREG} .

Now approximate confidence intervals for \bar{y}_N could be constructed based on a standard normal approximation to the distribution of $(\bar{y}_{\text{AREG}} - \bar{y}_N) / \{\hat{V}_g\}^{1/2}$. The justification of this procedure, from both a design and a model point of view, is asymptotic and the question of its appropriateness for particular finite samples must be addressed. One possibility is to compare

a set of confidence intervals obtained by this procedure to a set of purely model based intervals based on a further assumption of normality of errors and a t -statistic. If the two sets of intervals are wildly different there may be reason to doubt the validity of the jointly model and design based intervals, but more work is needed before this question can be answered satisfactorily.

An alternative approach to variance estimation in this framework is given by Binder (1983). The design variance of h^* as an estimator of g^* at $\hat{\gamma}_N$ could be estimated using standard design based techniques substituting $\hat{\gamma}_s$ for $\hat{\gamma}_N$, and then the variance of $\hat{\gamma}_s$ as an estimator of $\hat{\gamma}_N$ would be derived from a Taylor linearization of h^* about $\hat{\gamma}_N$. Taylor linearization could again be used to derive an estimator of the variance of a function of $\hat{\gamma}_s$ as an estimator of the same function of $\hat{\gamma}_N$.

4. AREAS FOR FURTHER RESEARCH

We have seen how the approach described here could be used for the estimation of finite population means or, more generally, for functions of linear regression parameters. It is natural to wonder whether and how the approach may be adapted to the estimation of other types of finite population parameters such as distribution functions and quantiles or to estimation for small areas.

Consider the special case of estimation of a distribution function at one point. There are two possible approaches to incorporate covariate information into a model. The first is to model the probability explicitly as a function of the covariates, an example is the logistic model. A second approach, which is common in the context of estimating a distribution function, as in Chambers and Dunstan (1986), Rao, Kovar and Mantel (1990), and others, is to model the residuals from a regression of the observed variable onto the covariables as being independent and identically distributed from some unknown distribution. The present approach requires that the parameter of interest be a function of the finite population parameter. Can this approach be adapted for the estimation of distribution functions or quantiles?

Another important problem in survey sampling is small area estimation, that is estimation of totals, means or proportions for subsets of the finite population. A good review is given in Platek, Rao, Särndal and Singh (1987). An obvious adaptation of the approach of Section 1 is to apply it separately within each domain of interest, what might be described as post-stratified generalized regression estimation. Note that this approach would require the totals of the covariates for each domain of interest. A very common approach in small area estimation is to borrow strength across areas via a model relating small areas to each other and to some covariates. A good review is given in Singh, Mantel and Thomas (1991). A very fruitful approach has been the empirical Bayes estimation based on random effects models which was introduced by Fay and Herriot (1979). Liang and Waclawiw (1990) discuss estimating functions for empirical Bayes models. Can the idea of modelling to borrow strength across small areas be formulated in such a way that the parameters of interest become functions of a population parameter?

ACKNOWLEDGEMENT

I am grateful to a referee and to the editor for helpful comments and suggestions.

REFERENCES

- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GODAMBE, V.P., and THOMPSON, M.E. (1984). Robust estimation through estimating equations. *Biometrika*, 71, 115-125.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- LIANG, K.-Y., and WACLAWIW, M.A. (1990). Extension of the Stein Estimating Procedure through the use of estimating functions. *Journal of the American Statistical Association*, 85, 435-440.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (Eds.) (1987). *Small Area Statistics An International Symposium*. New York: Wiley.
- SÄRNDAL, C.-E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. *Bulletin of the International Statistical Institute*, 49, 494-513.
- SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning Inference*, 7, 155-170.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1991). Time series generalizations of Fay-Herriot estimation for small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Maximum Likelihood Estimation from Complex Sample Surveys

ABBA M. KRIEGER and DANNY PFEFFERMANN¹

ABSTRACT

Maximum likelihood estimation from complex sample data requires additional modeling due to the information in the sample selection. Alternatively, pseudo maximum likelihood methods that consist of maximizing estimates of the census score function can be applied. In this article we review some of the approaches considered in the literature and compare them with a new approach derived from the ideas of 'weighted distributions'. The focus of the comparisons is on situations where some or all of the design variables are unknown or misspecified. The results obtained for the new method are encouraging, but the study is limited so far to simple situations.

KEY WORDS: Design adjusted estimators; Ignorable and informative designs; Pseudo likelihood; Weighted distributions.

1. INTRODUCTION

Survey data are often used for analytic inference about model parameters such as means, regression coefficients, cell probabilities *etc.* The models pertain to the population data and are therefore referred to as the census models. The problem in applying 'classical' maximum likelihood methods to survey data is that the model holding for the sample can be very different from the model holding for the population due to sample selection effects.

In order to illustrate the problem and some of the solutions proposed in the literature, consider the following simple example. A population U is made up of N units labelled $\{1, \dots, N\}$. Associated with unit i is a vector (Y_i, Z_i) of independent measurements drawn from a bivariate normal distribution with mean $\mu' = (\mu_Y, \mu_Z)$ and variance-covariance $(V - C)$ matrix

$$\Sigma = \begin{bmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{bmatrix}.$$

The values (y_i, z_i) are observed for a sample s of $n \ll N$ units selected by a probability sampling scheme. It is desirable to estimate μ_Y and σ_Y^2 . We consider three cases distinguished by the selection process and data availability.

Case A – The sample is selected by simple random sampling with replacement and only the values $\{(y_i, z_i), i \in s\}$ are known. Denoting the sample labels as $\{1, \dots, n\}$, we have that $Y_1, \dots, Y_n \underset{\text{ind}}{\sim} N(\mu_Y, \sigma_Y^2)$ yielding

$$\hat{\mu}_Y = \bar{y}_s = \sum_{i=1}^n y_i / n; \hat{\sigma}_Y^2 = \sum_{i=1}^n (y_i - \bar{y}_s)^2 / n = s_y^2 \quad (1.1)$$

as the MLE of μ_Y and σ_Y^2 . Clearly $E_M(\hat{\mu}_Y) = \mu_Y$ and $E_M\{[n/(n-1)]\hat{\sigma}_Y^2\} = \sigma_Y^2$ where $E_M\{\cdot\}$ defines the expectation under the model, with the sample units held fixed.

¹ Abba M. Krieger, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104. Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905.

Case B – The sample is selected with probabilities proportional to z_i with replacement such that at each draw $k = 1, \dots, n$, $P_i = P(i \in s) = z_i / \sum_{j=1}^N z_j$. The data known to the analyst are $\{y_i, z_i, i \in s\}$ and $\{z_{n+1}, \dots, z_N\}$. Suppose that $\text{Corr}(Y, Z) > 0$. This implies that $P(Y_i > \mu_Y \mid i \in s) > 1/2$ since the sampling scheme tends to select units with large values of Z and hence large values of Y . Clearly, the estimators defined in (1.1) are no longer MLE in this case.

The situation just described corresponds to the ‘classical’ example of missing data often analyzed in the literature (Anderson 1957). The MLE of μ_Y and σ_Y^2 are now

$$\hat{\mu}_Y = \bar{y}_s + b(\bar{Z} - \bar{z}_s); \hat{\sigma}_Y^2 = s_Y^2 + b^2(S_Z^2 - s_Z^2), \quad (1.2)$$

where $\bar{Z} = \sum_{i=1}^N z_i / N$, $\bar{z}_s = \sum_{i=1}^n z_i / n$, $b = \sum_{i=1}^n (y_i - \bar{y}_s)(z_i - \bar{z}_s) / \sum_{i=1}^n (z_i - \bar{z}_s)^2$, $S_Z^2 = \sum_{i=1}^N (z_i - \bar{Z})^2 / N$ and $s_Z^2 = \sum_{i=1}^n (z_i - \bar{z}_s)^2 / n$. Notice that the effect of the sample selection can be dealt with in this case by modeling the joint distribution of the response variable Y and the design variable Z . The sample selection process is then **ignorable** (see section 2.1).

Case C – Same as Case B but only the sample values $\{(y_i, z_i), i \in s\}$ and the sample selection probabilities $\{P_i, i \in s\}$ are known. Even though the values of z_i , $i = 1, \dots, N$ are known at the sampling stage, it is often the case that information on the design variables or the inclusion probabilities for units outside the sample is not included in the files released to analysts performing secondary analysis.

The estimators defined by (1.2) are no longer operational in this case since the population mean and variance of Z are unknown. For large populations, however, such that $\bar{Z} \approx \text{constant}$, an approximate MLE estimator of μ_Y is obtained as $\mu_Y^* = \bar{y}_s + b^*(1/N - \bar{P}_s)$ where $\bar{P}_s = \sum_{i=1}^n P_i / n$ and $b^* = \sum_{i=1}^n (y_i - \bar{y}_s)(P_i - \bar{P}_s) / \sum_{i=1}^n (P_i - \bar{P}_s)^2$. The rationale for μ_Y^* is that $P_i = Z_i / N\bar{Z}$ so that for $\bar{Z} = \text{constant}$, (Y_i, P_i) is bivariate normal with $\bar{P} = \sum_{i=1}^N P_i / N = 1/N$. This estimator is an example of using the sample selection probabilities as surrogates for the design variables when information on the latter is incomplete, as recommended in Rubin (1985).

A possible way to obtain approximate MLE under Case C is to follow what is known in the literature as the pseudo likelihood approach. We describe the approach in more detail in section 2, but it basically consists of maximizing a design consistent estimator of the census score function, that is, the score function that would have been obtained in the case of a census. The latter is unaffected by the design. Application of this approach yields, under Case C the estimators

$$\bar{\mu}_Y = \bar{y}_{ps} = \sum_{i=1}^n w_i^* y_i / \sum_{i=1}^n w_i^*; \bar{\sigma}_Y^2 = s_p^2 = \sum_{i=1}^n w_i^* (y_i - \bar{y}_{ps})^2 / \sum_{i=1}^n w_i^*, \quad (1.3)$$

where $w_i^* = (1/nP_i)$. Since \bar{y}_{ps} and s_p^2 are design consistent for $\bar{Y} = \sum_{i=1}^N y_i / N$ and $S_Y^2 = \sum_{i=1}^N (y_i - \bar{Y})^2 / N$ respectively, they are also consistent for μ_Y and σ_Y^2 in the sense that $\text{plim}_{n \rightarrow \infty, N \rightarrow \infty} (\bar{y}_{ps}, s_p^2) = (\mu_Y, \sigma_Y^2)$.

In this article we discuss a different approach for maximum likelihood estimation that is operational in principle even when the only information available to the analyst is the sample data. The method is derived from the theory of weighted distributions (Rao 1965, 1985, Patil and Rao 1978) and it utilizes the sample selection probabilities. The method is illustrated for the case of normal distributions with two different sampling designs and is shown to perform well in these cases. Another apparent advantage of the proposed approach emerging from the empirical study is that it is not very sensitive to misspecification of the design variables.

In section 2 we review the different approaches for MLE from survey data considered in the literature. Section 3 outlines the basic steps of the new approach. The empirical study is described and summarized in section 4. Section 5 contains concluding remarks.

2. REVIEW OF APPROACHES CONSIDERED IN THE LITERATURE

In this section we review briefly the approaches considered in the literature for MLE or approximate MLE from survey data. To better understand the complexity of the problem, we first discuss the notion of **ignorable sampling designs**. For a more detailed review of maximum likelihood and other approaches for analytic inferences from sample surveys see Pfeffermann (1993).

2.1 Ignorable and Informative Sampling Designs

Let $\underline{Z}' = (Z_1, \dots, Z_K)$ represent K design (auxiliary) variables used for designing the survey and denote by $Z = (z_1, \dots, z_N)'$ the $N \times K$ matrix of measurements on \underline{Z} so that z_i is the vector associated with unit i . The design variables may include strata indicator variables and quantitative measurements of cluster and unit characteristics. Let $\underline{Y}' = (Y_1, \dots, Y_p)$ represent the survey response variables. We assume for convenience that \underline{Y} is separate from \underline{Z} although as we mention below and consider in the empirical study, the sample selection probabilities may depend on the Y -values directly. The matrix $Y = (y_1, \dots, y_N)$ of the response variables values can be decomposed as $Y = [Y_s, Y_{\bar{s}}]$ where $Y_s = \{y_{is}, i \in s\}$ and $Y_{\bar{s}} = \{y_{i\bar{s}}, i \notin s\}$. Let $\underline{I} = (I_1, \dots, I_N)'$ be a vector of sample inclusion indicators such that $I_i = 1$ for $i \in s$ and $I_i = 0$ otherwise.

The basic problem of MLE from complex survey data, as illustrated in the introduction, is that in general, $f(Y_s; \underline{\lambda}^*) \neq \int f(Y_s, Y_{\bar{s}} | Z; \underline{\theta}_1) dY_{\bar{s}}$ where the symbol $f(\cdot; \cdot)$ defines probability density functions (pdf). As further illustrated in the introduction, this problem can sometimes be resolved by modeling the joint distribution of Y and Z . Thus, suppose that the values of \underline{Z} are known for every unit in the population and that \underline{Y} is observed for only the sample units. The joint pdf of all the available data can be written as

$$f(Y_s, \underline{I}; Z; \underline{\theta}, \underline{\phi}, \underline{\rho}) = \int f(Y_s, Y_{\bar{s}} | Z; \underline{\theta}_1) P(\underline{I} | Y, Z; \underline{\rho}_1) g(Z; \underline{\phi}) dY_{\bar{s}}. \quad (2.1)$$

Ignoring the sampling selection in the inference process implies that inference is based on the joint distribution of Y_s and Z , that is, the probability $P(\underline{I} | Y, Z; \underline{\rho}_1)$ on the right hand side of (2.1) is ignored. Hence the inference is based on

$$f(Y_s, Z; \underline{\theta}, \underline{\phi}) = \int f(Y_s, Y_{\bar{s}} | Z; \underline{\theta}_1) g(Z; \underline{\phi}) dY_{\bar{s}}. \quad (2.2)$$

The sample selection is said to be ignorable when inference based on (2.1) is equivalent to inference based on (2.2). This is clearly the case for sampling designs that depend only on the design variables \underline{Z} , since in this case $P(\underline{I} | Y, Z; \underline{\rho}_1) = P(\underline{I} | Z; \underline{\rho}_1)$. The exact conditions for the ignorability of the sample selection process are defined and illustrated in the articles by Rubin (1976), Little (1982) and Sugden and Smith (1984).

The complications of MLE from complex survey data based on (2.1) or (2.2) are now apparent. First and foremost, it requires that all the relevant design variables be identified and known at the population level. As often argued in the literature, (see Pfeffermann 1993 for references), this is not necessarily the case. Secondly, it requires that the sample selection is ignorable in the sense discussed above or alternatively that the probabilities $P(\underline{I} | Y, Z; \underline{\rho})$ be modeled and included in the likelihood. Finally, the use of MLE requires the specification of the joint pdf $f(Y, Z; \underline{\theta}, \underline{\phi}) = f(Y | Z; \underline{\theta}_1) g(Z; \underline{\phi})$.

2.2 Exact MLE Based on Factorization of the Likelihood

Factoring the likelihood in the case of multivariate normal data was first suggested by Anderson (1957). The factorization is possible when the observed data have a nested pattern, that is, the set of survey variables X_1, \dots, X_p can be arranged such that X_j is observed for all units where X_{j+1} is observed, $j = 1, \dots, (p - 1)$. Extensions to other distributions and more general data patterns are given in Rubin (1974). Holt, Smith and Winter (1980) apply the ideas to MLE of regression coefficients from complex survey data.

Suppose that the sample selection is ignorable so that inference can be based on the joint distribution $f(Y_s, Z; \underline{\theta}, \underline{\phi}) = f(Y_s | Z; \underline{\theta}_1) g(Z; \underline{\phi})$. The likelihood can be factored accordingly as

$$L(\underline{\theta}, \underline{\phi}; Y_s, Z) = L(\underline{\theta}_1; Y_s | Z) L(\underline{\phi}; Z). \quad (2.3)$$

Assuming that the parameters $\underline{\theta}_1$ and $\underline{\phi}$ are distinct in the sense of Rubin (1976), MLE of $\underline{\theta}_1$ and $\underline{\phi}$ can be calculated independently from the two components.

Application of (2.3) to the case where (Y'_i, Z'_i) are multivariate normal yields the following MLE for $\underline{\mu}_Y = E(\underline{Y})$ and $\underline{\Sigma}_Y = V(\underline{Y})$ (Anderson 1957).

$$\hat{\underline{\mu}}_Y = \bar{y}_s + \hat{\underline{\beta}}(\bar{z} - \bar{z}_s); \quad \hat{\underline{\Sigma}}_Y = s_{YY} + \hat{\underline{B}}[S_{ZZ} - s_{ZZ}]\hat{\underline{B}}', \quad (2.4)$$

where $(\bar{y}_s, \bar{z}_s) = \sum_{i=1}^n (y_i, z_i)/n$, $\bar{z} = \sum_{i=1}^N z_i/N$, $S_{ZZ} = \sum_{i=1}^N (z_i - \bar{z})(z_i - \bar{z})'/N$, $s_{ZZ} = \sum_{i=1}^n (z_i - \bar{z}_s)(z_i - \bar{z}_s)'/n$ and $\hat{\underline{B}} = \sum_{i=1}^n (y_i - \bar{y}_s)(z_i - \bar{z}_s)' s_{ZZ}^{-1}/n$.

The MLE of the coefficient matrix B_{12} of the multivariate regression of Y_1 on Y_2 where $\underline{Y}' = (Y'_1, Y'_2)$ is obtained straightforwardly from (2.4). Thus, if

$$\underline{\Sigma}_Y = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where

$$\Sigma_{ij} = \text{COV}[(Y'_i, Y'_j)'], \quad i, j = 1, 2, \quad B_{12} = \Sigma_{12} \Sigma_{22}^{-1} \quad \text{and} \quad \hat{B}_{12} = \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1}.$$

For the explicit expression of \hat{B}_{12} see Holt, Smith and Winter (1980).

2.3 Design Adjusted Estimators (DAE)

Assume that the sample selection mechanism is ignorable. Let $\ell_N(\theta; Y)$ denote the log likelihood for θ that would be obtained in the case of a census. Denote by $h_N(Y | Z, Y_s; \underline{\theta}_1)$ the conditional distribution of Y given Z and Y_s and let $E_{h_N}(\cdot | Z, Y_s)$ define the expectation operator under h_N . The DAE $\hat{\underline{\theta}}_{ND}$ of $\underline{\theta}$ as proposed by Chambers (1986) is defined as

$$E_{h_N}[-\ell_N(\hat{\underline{\theta}}_{ND}) | Z, Y_s] = \min\{E_{h_N}[-\ell_N(\underline{\theta}) | Z, Y_s]; \underline{\theta} \in \Theta\}. \quad (2.5)$$

Notice that the expectation $E_{ND}(\theta) = E_{h_N}[\ell_N(\theta) | Z, Y_s]$ depends on the vector parameter $\underline{\theta}_1$ of the conditional distribution $f(Y | Z; \underline{\theta}_1)$. The estimator $\hat{\underline{\theta}}_{ND}$ of (2.5) is computed by substituting $\hat{\underline{\theta}}_1$ for $\underline{\theta}_1$ where $\hat{\underline{\theta}}_1$ is the MLE of $\underline{\theta}_1$ obtained from the data (Y_s, Z) .

Simple algebra shows that for the multivariate normal model considered in section 2.2, the DAE of μ_Y and Σ_Y are the same as the MLE defined by (2.4). A possible advantage of this approach, however, is that it can be applied to other loss functions.

2.4 The Pseudo Likelihood Approach

The prominent feature of this approach is that it utilizes the sample selection probabilities to estimate the census likelihood equations. The estimated equations are then maximized with respect to the vector parameter of interest. No information on the values of the design variables is needed, although as illustrated in the empirical study, knowledge of these values at the population level can be used to improve the efficiency of the estimators.

Suppose that the population values Y_i are independent draws from a common distribution $f(Y; \theta)$ and let $\ell_N(\theta; Y) = \sum_{i=1}^N \log f(Y_i; \theta)$ define the census log likelihood function. Under some regularity conditions, the MLE, $\hat{\theta}$, solves the equations

$$U(\theta) = d\ell_N(\theta; Y)/d\theta = \sum_{i=1}^N u(\theta; y_i) = 0, \quad (2.6)$$

where “ d ” defines the derivative operator and $u(\theta; y_i) = d \log f(Y_i; \theta)/d\theta$. The pseudo MLE of θ is defined as the solution of $\tilde{U}(\theta) = 0$ where $\tilde{U}(\theta)$ is a design consistent estimator of $U(\theta)$ in the sense that $\text{plim}_{n \rightarrow \infty, N \rightarrow \infty} [\tilde{U}(\theta) - U(\theta)] = 0$ for all $\theta \in \Theta$. The commonly used estimator of $U(\theta)$ is the Horvitz-Thompson (1952) estimator so that the pseudo MLE of θ is the solution of $\tilde{U}(\theta) = \sum_{i=1}^n w_i^* u(\theta; y_i) = 0$ where for selection without replacement $w_i^* = [1/P(i \in s)]$ and for selection with replacement $w_i^* = (1/nP_i)$.

For the multivariate normal model, the pseudo MLE of μ_Y and Σ_Y are

$$\tilde{\mu}_Y = \sum_{i=1}^n w_i^* y_i / \sum_{i=1}^n w_i^*; \quad \tilde{\Sigma}_Y = \sum_{i=1}^n w_i^* (y_i - \tilde{\mu}_Y)(y_i - \tilde{\mu}_Y)' / \sum_{i=1}^n w_i^*. \quad (2.7)$$

The pseudo MLE of the matrix coefficients B_{12} is obtained as $\tilde{B}_{12} = \tilde{\Sigma}_{12} \tilde{\Sigma}_{22}^{-1}$.

Various examples for the use of this approach under different models can be found in Skinner *et al.* (1989). See also Binder (1983), Chambless and Boyle (1985), Roberts, Rao and Kumar (1987) and Pfeiffermann (1988).

Information on auxiliary design variables known at the population level can be used to improve the efficiency of the design estimators of $U(\theta)$. The “probability weighted MLE” as proposed by Nathan and Holt (1980) and by Smith and Holmes (Skinner *et al.* 1989, Ch. 8) are examples of the use of the population values of the design variables. The estimators have the same structure as the exact MLE derived from (2.4) but with unweighted sample statistics replaced by weighted statistics. For example, (\bar{y}_s, \bar{z}_s) in (2.4) are replaced by $\sum_{i=1}^n w_i^* (y_i, z_i) / \sum_{i=1}^n w_i^*$, with similar substitutions for the other expressions.

An important property of pseudo MLE is that they are in general design consistent for the population quantities that would be obtained by solving the corresponding census likelihood equations, irrespective of whether the model is correct and/or whether the sampling design is informative. See Pfeiffermann (1993) for the implications of this property with references to other studies. Other theoretical properties of pseudo MLE are studied by Godambe and Thompson (1986).

3. MLE DERIVED FROM WEIGHTED DISTRIBUTIONS

3.1 General Formulation

The weighted pdf of a random variable X^w is defined as

$$f^w(x) = w(x)f(x)/w, \quad (3.1)$$

where $f(x)$ is the unweighted pdf and $w = \int w(x)f(x)dx = E[w(X)]$ is the normalizing factor making the total probability equal to unity. Situations leading to weighted distributions occur when realizations x from $f(x)$ are observed and recorded with differential probabilities $w(x)$. The expectation w is then the probability of recording an observation and $f^w(x)$ is the pdf of the resulting random variable X^w .

The concept of weighted distributions was introduced by Rao (1965). Patil and Rao (1978) discuss various practical situations that give rise to pdf's of the form (3.1). One special case that occurs in many applications is when $w(x) = |x|$ where $|x|$ is some measure of the size of x . The pdf obtained in this case is called 'size biased' or 'length biased'. The properties of that distribution under a variety of densities $f(x)$ are examined in Cox (1969) and Patil and Rao (1978). Estimation of weighted distributions is considered by Vardi (1982).

How can the concept of weighted distributions be utilized for analytic inference from complex samples? Consider as before a finite population $U = \{1, \dots, N\}$ with random measurements $X(i) = \underline{x}_i' = (y_i', z_i')$ generated independently from a common pdf $h(\underline{x}; \underline{\delta}) = f(y_i | z_i; \theta_1) g(z_i; \phi)$. Suppose that unit i is sampled with probability $w(\underline{x}_i; \underline{\alpha})$ that depends on the measurements \underline{x}_i and possibly also on an unknown vector parameter $\underline{\alpha}$. Denote by X_i^w the measurements recorded for unit i 's. The pdf of X_i^w is then

$$\begin{aligned} h^w(\underline{x}_i; \underline{\alpha}, \underline{\delta}) &= f(\underline{x}_i | i \in s) = P[i \in s | X(i) = \underline{x}_i] h(\underline{x}_i; \underline{\delta}) / P(i \in s) \\ &= w(\underline{x}_i; \underline{\alpha}) h(\underline{x}_i; \underline{\delta}) / \int w(\underline{x}; \underline{\alpha}) h(\underline{x}; \underline{\delta}) d\underline{x}. \end{aligned} \quad (3.2)$$

Analytic inference focuses on the vector parameter $\underline{\delta}$ or functions thereof as the target parameters. Let $s = \{1, \dots, n\}$ define a sample of fixed size $n \ll N$ selected with replacement such that at each draw $k = 1, \dots, n$, $P(j \in s) = w(\underline{x}_j; \underline{\alpha})$, $j = 1, \dots, N$. The joint pdf of $\{X_i^w, i = 1, \dots, n\}$ is then $\prod_{i=1}^n h^w(\underline{x}_i; \underline{\alpha}, \underline{\delta})$ so that the likelihood is

$$L(\underline{\delta}; X_s, s) = \text{const} \times \prod_{i=1}^n h(\underline{x}_i; \underline{\delta}) / [\int w(\underline{x}; \underline{\alpha}) h(\underline{x}; \underline{\delta}) d\underline{x}]^n, \quad (3.3)$$

where $X_s' = [\underline{x}_1, \dots, \underline{x}_n]$. The likelihood (3.3) has the following properties:

- (1) It is defined in terms of the vector parameter $\underline{\delta}$. This has an advantage over the use of the factorized likelihood (2.3) where $\underline{\delta}$ does not enter the likelihood directly.
- (2) It is a function of the selection probabilities $w(\underline{x}_i; \underline{\alpha})$ that enter into the denominator.
- (3) The likelihood relates to the conditional distribution of the sample data given the units in the sample. This is different from the likelihood derived from the pdf in (2.1) which is the joint pdf of the sample data and the vector \underline{l} of sample indicators. An example of the use of the latter pdf in conjunction with weighted distributions for MLE is given in Godambe and Rajarshi (1989).

- (4) The use of the likelihood (3.3) requires a definition of the joint pdf $h(x; \delta)$ holding in the population and a specification of the relationship between the sample selection probabilities and the variables observed for the sample. The need to define the population pdf is common to all of the approaches for MLE proposed in the literature. The specification of the functions $w(x)$ is unique to the present approach. This step can be carried out however by modeling the empirical relationship in the sample between the selection probabilities and the observed measurements. Having identified a suitable model, the probabilities $w(x, \alpha)$ can be estimated from the sample and the estimates can be substituted into the likelihood. In what follows we consider two examples which are analyzed empirically in section 4.

3.2 Examples

We assume the model considered in section 2 in which $X'_i = (Y'_i, Z'_i)$ are independent realizations from a multivariate normal distribution with mean $\mu'_x = (\mu'_Y, \mu'_Z)$ and $V - C$ matrix

$$\Sigma_{XX} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}. \quad (3.4)$$

Consider the following sampling designs:

D1 - PPS selection with replacement: Let $T_i = \alpha'_1 Y_i + \alpha'_2 Z_i$ define a single design variable and suppose that the sample is selected with probabilities proportional to the T -values such that at each draw $k = 1, \dots, n$, $P(i \in s) = t_i / N\bar{T}$, $i = 1, \dots, N$ where $\bar{T} = \sum_{j=1}^N t_j / N$. We assume that N is large enough so that the difference between \bar{T} and $\mu_T = E(T)$ can be ignored. The coefficients $\alpha = (\alpha'_1, \alpha'_2)$ are fixed. In special cases $\alpha_1 = 0$ hence T is a function of only the auxiliary design variables Z or $\alpha_2 = 0$ in which case T is only a function of the response variables Y . Suppose as before that it is desirable to estimate the mean μ_Y and the $V - C$ matrix Σ_{YY} or functions thereof.

When $\alpha_1 = 0$ and T is known for every unit in the population, one can estimate the unknown parameters using the factorization (2.3). The corresponding MLE are given in (2.4) with Z replaced by T . Suppose however that the only information available to the analyst is the sample values $x'_i = (y'_i, z'_i)$, $i = 1, \dots, n$ and the sample selection probabilities $P_i = t_i / N\bar{T}$. Under the assumption $\bar{T} = \mu_T$, the likelihood for $[\mu_X, \Sigma_{XX}]$ can be written using (3.3) as

$$L(\mu_X, \Sigma_{XX}; X_s, s) = \prod_{i=1}^n (\alpha'_i x_i) \phi(x_i; \mu_X, \Sigma_{XX}) / (\alpha'_1 \mu_Y + \alpha'_2 \mu_Z)^n, \quad (3.5)$$

where $\phi(x; \mu_X, \Sigma_{XX})$ is the normal pdf with mean μ_X and $V - C$ matrix Σ_{XX} . The likelihood in (3.5) is a function also of the unknown vector coefficients α . However, the values of α can actually be found up to a constant c (which cancels out in the likelihood) by regressing the sample selection probabilities P_i against α .

In the simulation study described in section 4, we consider the case where not all the design variables are known even for the sample units. Thus, suppose that $Z'_i = (Z_{1i}, Z_{2i})$ and that the data available to the analyst consist of the selection probabilities P_i , $i = 1, \dots, n$ and the observations $\{x_i^{*'} = (y'_i, z_{1i}), i = 1, \dots, n\}$. The likelihood (3.3) is now

$$L(\underline{\mu}_x^*, \sum_{XX}^*; X_s^*, s) = \prod_{i=1}^n w(\underline{x}_i^*) \phi(\underline{x}_i^*; \underline{\mu}_x^*, \sum_{XX}^*) / (w^*)^n, \quad (3.6)$$

where $w(\underline{x}_i^*)$ are the selection probabilities expressed as functions of \underline{x}_i^* . Clearly, the probabilities $w(\underline{x}_i^*)$ are not fully determined by the values \underline{x}_i^* unless $\alpha_{22} = 0$. Assuming normality

$$w(\underline{x}_i, \underline{\alpha}) = \alpha_0^* + \alpha_1^{*'} \underline{y}_i + \alpha_2^{*'} z_{1i} + \epsilon_i, \quad (3.7)$$

where $\{\epsilon_i\}$ is white noise. Thus, the likelihood (3.6) can be approximated by substituting $w^*(\underline{x}_i^*) = \alpha_0^* + \alpha_1^{*'} \underline{y}_i + \alpha_2^{*'} z_{1i}$ for $w(\underline{x}_i^*)$. The values of $\underline{\alpha}^* = (\alpha_0^*, \alpha_1^{*'}, \alpha_2^{*'})'$ can be estimated from the regression (3.7) and then substituted into the likelihood.

D2 – Stratified sampling with T as the stratification variable: Suppose that the population U is divided into L strata U_1, \dots, U_L of sizes N_1, \dots, N_L , $\sum_{h=1}^L N_h = N$, based on the ascending values of T . Consider a simple random stratified sample of size $n = \sum_{h=1}^L n_h$ selected without replacement with fixed sample sizes $\{n_h\}$. The weighted pdf of X_i^w , the measurements recorded for unit $i \in s$ is in this case [compare with (3.2)]

$$h^w(\underline{x}_i; \underline{\alpha}, \underline{\delta}) = f(\underline{x}_i \mid i \in s) = \begin{cases} P_1 h(\underline{x}_i; \underline{\delta}) / w & \text{if } t_i \leq t^{(1)} \\ P_2 h(\underline{x}_i; \underline{\delta}) / w & \text{if } t^{(1)} \leq t_i \leq t^{(2)} \\ \vdots & \vdots \\ P_L h(\underline{x}_i; \underline{\delta}) / w & \text{if } t^{(L-1)} \leq t_i \end{cases} \quad (3.8)$$

where $P_h = (n_h/N_h)$ and for $\{N_h\}$ sufficiently large, the probability $w = P(i \in s)$ can be closely approximated as

$$w = P(i \in s) \approx P_1 \int_{-\infty}^{t^{(1)}} \phi(t) dt + \sum_{h=2}^{L-1} P_h \int_{t^{(h-1)}}^{t^{(h)}} \phi(t) dt + P_L \int_{t^{(L-1)}}^{\infty} \phi(t) dt, \quad (3.9)$$

where $\phi(t)$ denotes the normal pdf of T .

Suppose that the strata are large enough so that selection within the strata can be considered as independent. Define $\mu_T = E(T) = \underline{\alpha}' \underline{\mu}_X$, $\sigma_T^2 = \text{Var}(T) = \underline{\alpha}' \sum_{XX} \underline{\alpha}$ and let $\Phi_h = \int_{-\infty}^{t^{(h)}} \phi(t) dt$. For given boundaries $\{t^{(h)}\}$ and the vector coefficients $\underline{\alpha}$, the likelihood for $\underline{\delta}$ can be written as

$$L(\underline{\delta}; X_s, s) = \text{const} \times \prod_{i=1}^n h(\underline{x}_i; \underline{\delta}) \prod_{h=1}^L P_h^{n_h} / \left\{ P_1 \Phi_1 + \sum_{h=2}^{L-1} P_h [\Phi_h - \Phi_{h-1}] + P_L [1 - \Phi_{L-1}] \right\}^n. \quad (3.10)$$

Hausman and Wise (1981) use a variant of the likelihood (3.10) for estimating the vector of regression coefficients in a situation where the strata boundaries are determined by the values of the dependent variable. They assume that the strata boundaries are known, but allow the selection probabilities within the strata to be unknown in which case they are included in the set of unknown parameters with respect to which the likelihood is maximized.

In many practical situations, the strata boundaries are unknown and have to be estimated from the sample data. When the data include the values $\{t_i, i = 1, \dots, n\}$, the vector α can be estimated from the regression of t_i on x_i , as in the PPS example discussed before. Furthermore, if $(t_{(1)} \leq \dots \leq t_{(n)})$ are the ordered values of the t_i 's, the strata boundaries can be estimated as, $t^{(1)} = 1/2(t_{(n_1)} + t_{(n_1+1)}) \dots t^{(L-1)} = 1/2(t_{(n^*)} + t_{(n^*+1)})$ where $n^* = \sum_{h=1}^{L-1} n_h$. Substituting these estimates into (3.10) yields an approximation to the likelihood which can then be maximized as a function of $\hat{\delta}$.

The situation is more complicated when the values t_i are unknown even for units in the sample. In the simulation study we attempt to deal with this problem by predicting t_i using Fisher's Linear Discriminant Function, that is, specifying the vector coefficients $\hat{\alpha}$ to be such that it maximizes the ratio of the between groups sum of squares to the within groups sum of squares of linear combinations $\hat{\alpha}' X_i$. The groups are the strata. Once the predictors $\hat{t}_i = \hat{\alpha}' x_i$ are formed, the strata boundaries are estimated as in the previous case but with \hat{t}_i instead of t_i . Also, $\hat{\mu}_T = \hat{\alpha}' \mu_X$ and $\hat{\sigma}_T^2 = \hat{\alpha}' \sum_{XX} \hat{\alpha}$. Substituting these estimators in (3.10) yields an approximation to the likelihood which can be maximized with respect to $\hat{\delta}$.

As in the PPS example, the likelihood (3.10) can be modified to the case where only some of the design variables are known or observed. Maximization of the modified likelihood is carried out following the same steps as above.

4. SIMULATION RESULTS

4.1 General

In order to illustrate and compare the performance of the various MLE procedures described in this paper, we ran a small simulation study which consists of two stages. In the first stage we generated a single finite population of size $N = 8,000$ such that $x_i' = (y_{1i}, y_{2i}, z_{1i}, z_{2i})$, $i = 1, \dots, 8,000$ are multivariate normal. In the second stage we selected independent samples of size $n = 800$ using the two sampling schemes described in section 3.2 with two different definitions for the design variable. The number of samples selected in each case was 300. We computed the various estimators for each of the samples based on the available sample data and then computed the empirical bias and root mean square error (RMSE) over the selected samples. In order to study and compare the conditional properties of the estimators considered, we classified the 300 samples selected in each case into 10 groups, based on the ascending values of the sample mean of the design variable and computed the bias and RMSE within each of the groups. In what follows we describe the various stages in some more detail.

4.2 Generation of the Population Values and Sample Selection Schemes

Values of z_{1i} and z_{2i} were generated independently from a normal $(20, 10^2)$ distribution. Values y_{1i} were generated as $y_{1i} = z_{1i} + z_{2i} + \epsilon_{1i}$; $\epsilon_{1i} \sim N(0, 10^2)$. Values y_{2i} were generated as $y_{2i} = y_{1i} + 0.5z_{1i} + 0.5z_{2i} + \epsilon_{2i}$; $\epsilon_{2i} \sim N(0, 20^2)$.

We employed the two sampling schemes described in section 3.2 using two different definitions for the design size variable. (i) $t_i = 0.5(z_{1i} + z_{2i})$ and (ii) $t_i = 0.25(y_{1i} + y_{2i} + z_{1i} + z_{2i})$. Thus, selection based on the first design variable satisfies the ignorability conditions defined in section 2.1, provided that the data for (Z_1, Z_2) are known for the entire population. When these data are only known for the sample, the sampling design is ignorable only with respect to the conditional distribution $f(y_1, y_2 | z_1, z_2)$. When selection is based on the second design variable, the sampling design is informative.

For the stratified selection D2, we generated eight equal sized strata defined by the ascending values of the size variable. The sample sizes within the strata were such that they increase with increasing values of the t'_i s.

4.3 Estimators Considered

The parameters estimated in our study are the mean vector and the $V - C$ matrix of the marginal distribution of (Y_1, Y_2) . We consider seven different estimators for the design D1 and nine estimators for the design D2. See section 3.2 for description of the computations involved in the derivation of the various estimators.

DESIGN D1

- $ML(Z_1, Z_2)$ – The exact MLE for the case where the design is ignorable, (equation 2.4).
- $WML(Z_1, Z_2)$ – The estimators obtained from $ML(Z_1, Z_2)$ by replacing the unweighted sample statistics by probability weighted statistics (see the discussion below equation 2.7).
- $ML(Z_1)$ – Same as $ML(Z_1, Z_2)$ but with Z_1 as the only design variable so that $\underline{Z} = Z_1$.
- $WML(Z_1)$ – Same as $WML(Z_1, Z_2)$ but with Z_1 as the only design variable.
- CPL – The classical pseudo likelihood estimators (equations 2.7).
- $WDML(X^*)$ – The (weighted distribution) estimators obtained by maximization of the likelihood in (3.6).
- $WDML(X^*, Z_1)$ – The estimators obtained by maximizing the likelihood in (3.6) but with the mean and variance of Z_1 fixed at their population values.

DESIGN D2

The first 5 estimators are the same as the estimators for the design D1. The other 4 estimators are defined as follows:

- $WDML(X^*)$ – The estimators obtained by maximizing the likelihood (3.10) with the α^* – coefficients [(equation (3.7))] estimated by the linear discriminant function.
- $WDML(X^*, Z_1)$ – Same as $WDML(X^*)$ but with the mean and variance of Z_1 fixed at their population values.
- $WDML(X^*, t_s)$ – The estimators obtained by maximizing the likelihood (3.10) when the values $t_s = (t_1, \dots, t_n)$ are known for units in the sample.
- $WDML(X^*, t_s, Z_1)$ – Same as $WDML(X^*, t_s)$ but with the mean and variance of Z_1 fixed at their population values.

It should be emphasized that the estimators derived based on the weighted distributions are not really MLE because of the approximations involved in the maximization procedures as described in section 3.2 (see also comment 2 below).

Comments

- (1) The estimators we consider can be classified according to the sample and population data they use and according to whether the design variables are correctly specified and the ignorability conditions are met. Thus, the estimators $ML(Z_1, Z_2)$ and $WML(Z_1, Z_2)$ use the population values of Z_1 and Z_2 and the sample values of Y_1 and Y_2 . As mentioned in section 2.4 and further discussed in Pfeffermann (1993), the use of $WML(Z_1, Z_2)$ is to protect against possible model misspecifications or informative sampling schemes. The estimators $ML(Z_1)$, $WML(Z_1)$, $WDML(X^*, Z_1)$ and $WDML(X^*, t_s, Z_1)$ use the known population data for Z_1 but not the data for Z_2 even for the sample units. The use of these estimators corresponds to situations where the design variables are misspecified or the values of some of them are unknown. The estimator $WDML(X^*)$ uses only the sample information for Y_1 , Y_2 and Z_1 and the sample selection probabilities. The estimator $WDML(X^*, t_s)$ uses in addition the sampling values of the design variable. The estimator CPL uses only the sample values of Y_1 and Y_2 and the sample selection probabilities.
- (2) We maximized the likelihood derived from the weighted distributions using a quasi-Newton method in the subroutine library IMSL. The method employed requires partial derivatives of the likelihood with respect to each of the parameters as user supplied input. An issue that arose in the maximization is worth mentioning. It is easier to parameterize the likelihood in terms of Σ^{-1} where Σ is the covariance matrix among Y_1 , Y_2 and Z_1 . Furthermore, to insure that the six parameters that define Σ^{-1} are unconstrained, we use the elements of the upper triangular matrix B so that $B'B = \Sigma^{-1}$. Any choice of the values for B leads to a matrix Σ^{-1} that is positive semi-definite.

4.4 Results

We present the results obtained when estimating $\mu_1 = E(Y_1)$, $\sigma_1^2 = \text{Var}(Y_1)$ and B_{21} - the slope coefficient in the regression of Y_2 on Y_1 , as representative of the results obtained when estimating the other parameters. Tables 1-3 contain the RMSE of the various estimators as obtained for the two sampling schemes and the two choices of the design variable. RMSE's dominated by large biases are indicated by an asterisk.

The main results emerging from the tables (and from estimating the other model parameters) can be summarized as follows:

- (1) The estimator $ML(Z_1, Z_2)$ outperforms all of the other estimators when the ignorability conditions are met, but it is severely biased when the sampling design is informative. The estimator $WML(Z_1, Z_2)$ is essentially unbiased in all of the cases, but the use of the sampling weights increases the variance. Still, this estimator dominates in general the estimator CPL especially under the PPS design because of the use of the population values of (Z_1, Z_2) .
- (2) The estimator $ML(Z_1)$ is severely biased in almost all of the cases. Notice in particular the large biases in the case where $t_i = 0.5(z_{1i} + z_{2i})$, illustrating the sensitivity of the MLE's to the exact specification of the design variables. Like with $WML(Z_1, Z_2)$, the estimator $WML(Z_1)$ is unbiased, and for the PPS design it outperforms the estimator CPL.
- (3) The estimator CPL is unbiased in all of the cases. An interesting result emerging from the tables is that relative to the other estimators considered, it performs better in estimating the mean than in estimating variances and covariances. An intuitive explanation for this outcome is that in the latter case the sampling weights are used twice, thereby increasing the variance.

Table 1
RMSE of Estimators of μ_1 for Different Sampling Schemes and Design Variables
(True Mean: $\mu_1 = 40$)

Estimators	D1 – PPS Sampling		D2 – Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1, Z_2)$	0.43	1.86*	0.47	3.43*
$WML(Z_1, Z_2)$	0.43	0.57	0.50	0.52
$ML(Z_1)$	2.67*	4.38*	6.39*	8.32*
$WML(Z_1)$	0.58	0.90	0.62	0.58
$WDML(X^*, Z_1)$	0.56	0.63	1.51*	0.59
$WDML(X^*)$	0.80	0.90	3.59*	0.49
CPL	0.77	1.19	0.56	0.47
$WDML(X^*, t_s)$	–	–	0.74	0.43
$WDML(X^*, t_s, Z_1)$	–	–	0.74	0.57

* RMSE dominated by bias.

Table 2
RMSE of Estimators of σ_1^2 for Different Sampling Schemes and Design Variables
(True Variance: $\sigma_1^2 = 300$)

Estimators	D1 – PPS Sampling		D2 – Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1, Z_2)$	12.33	18.35*	16.00	29.00*
$WML(Z_1, Z_2)$	14.00	18.72	20.87	19.83
$ML(Z_1)$	24.32*	33.66*	35.16*	53.66*
$WML(Z_1)$	18.61	26.61	24.22	20.35
$WDML(X^*, Z_1)$	14.36	17.41	26.94*	15.49
$WDML(X^*)$	16.37	19.68	41.08*	15.34
CPL	21.11	29.06	24.19	20.18
$WDML(X^*, t_s)$	–	–	26.18*	15.46
$WDML(X^*, t_s, Z_1)$	–	–	25.70*	15.72

* RMSE dominated by bias.

Table 3
RMSE of Estimators of B_{21} for Different Sampling Schemes and Design Variables
(True Coefficient: $B_{21} = 1.33$)

Estimators	D1 – PPS Sampling		D2 – Stratified Sampling	
	$t_i = 0.5z_i$	$t_i = 0.25x_i$	$t_i = 0.5z_i$	$t_i = 0.25x_i$
$ML(Z_1, Z_2)$	0.043	0.069*	0.048	0.120*
$WML(Z_1, Z_2)$	0.054	0.060	0.068	0.066
$ML(Z_1)$	0.045	0.078*	0.056	0.134*
$WML(Z_1)$	0.055	0.062	0.069	0.065
$WDML(X^*, Z_1)$	0.043	0.047	0.049	0.045
$WDML(X^*)$	0.044	0.049	0.050	0.046
CPL	0.055	0.063	0.069	0.065
$WDML(X^*, t_{5s})$	–	–	0.048	0.045
$WDML(X^*, t_{5s}, Z_1)$	–	–	0.048	0.045

* RMSE dominated by bias.

- (4) For the PPS design, the estimators $WDML(X^*)$ and $WDML(X^*, Z_1)$ perform very well with $WDML(X^*)$ clearly dominating CPL and $WDML(X^*, Z_1)$ dominating $WML(Z_1)$. Interestingly, the estimator $WDML(X^*)$ performs in general better than the estimator $WML(Z_1)$ despite the use of less information. The fact that $WDML(X^*)$ outperforms CPL could be explained by the fact that it is more “model dependent”, although as discussed in section (2.4), one way of viewing CPL is as the estimator maximizing the design unbiased estimator of the likelihood equations holding in the population.
- (5) Next consider the stratified design. In the case were $t_i = 0.25x_i$, the picture is very similar to the PPS case with $WDML(X^*)$ dominating again both CPL and $WML(Z_1)$. Actually, there is little to choose in this case among the four estimators derived from the weighted distribution likelihood despite the use of different sample and population data by each estimator. When $t_i = 0.5z_i$, all of the four estimators are inferior to $WML(Z_1)$ and CPL although interestingly enough, not with respect to the estimation of the regression coefficient where they all perform very similar to the optimal $ML(Z_1, Z_2)$. The particularly poor performance of $WDML(X^*)$ (and to a much lesser extent of $WDML(X^*, Z_1)$) in estimating the mean and variance is mainly the result of incorrect specification of the strata boundaries and hence incorrect specification of the denominator of the likelihood (3.10). This problem can possibly be resolved by either including the strata boundaries and the α^* – coefficients relating the values t_i to the observed data (equation 3.7) as part of the unknown parameters in the likelihood (3.10), or by replacing the linear discriminant function by some other (nonlinear) function such as logistic regression. The latter approach has the advantage of reducing the number of parameters over which the likelihood has to be maximized, which can be crucial when the number of strata is large.

We considered so far the unconditional bias and RMSE of the estimators. As mentioned in section 4.1, we studied also conditional properties by computing the bias and RMSE's over samples with similar sample means of the design variable. The conclusions reached from that study are very similar to the conclusions stated before. Thus, estimators which are approximately unbiased unconditionally are also approximately conditionally unbiased and vice versa.

This result is somewhat surprising because it has often been illustrated in the literature that the CPL estimator, for example, has poor conditional properties. Possible explanations in our case are that the sample size considered is large or that the division of the sample into the ten groups was not sharp enough. Because of space limitations we omit the results illustrating conditional properties of the estimators.

5. CONCLUDING REMARKS

The results of the simulation study show that estimators obtained by maximizing the likelihood derived from weighted distributions are a favorable alternative to the pseudo likelihood estimators obtained by maximizing design consistent estimators of the census likelihood equations. The estimators perform particularly well in our study when using an informative sampling scheme for which the "classical" MLE can become severely biased. The use of these estimators requires, however, the modeling of the relationship between the sample selection probabilities and the observed sample data. As illustrated in the simulation study, failure to model or estimate the relationship correctly may introduce large biases.

The key question to the practical use of these estimators is therefore whether the model relating the sample selection probabilities to the observed response and design variables can be successfully identified from the sample data. It would seem that this question can only be answered by considering actual surveys that use common sampling designs. Other important questions related to the use of these estimators are the availability of reliable variance estimators so that accurate confidence intervals can be set and the protection against misspecification of the parent distribution of the response variables in the population. These two questions are common to other MLE procedures. We hope that the initial results of our study will encourage further research on these and other related questions.

ACKNOWLEDGMENT

The work of Danny Pfeffermann was supported by the Statistics Canada Research Fellowship Program.

REFERENCES

- ANDERSON, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52, 200-203.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- CHAMBERS, R.L. (1986). Design adjusted parameter estimation. *Journal of the Royal Statistical Society A*, 149, 161-173.
- CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.

- COX, D.R. (1969). Some sampling problems in technology. In: *New Developments in Survey Sampling*, (Eds. N. Johnson and H. Smith Jr.). New York: Wiley, 506-527.
- GODAMBE, V.P., and RAJARSHI, M.B. (1989). Optimal estimation for weighted distributions: semiparametric models. In *Statistical Data Analysis and Inference*, (Ed. Y. Dodge). Amsterdam: Elsevier Science, 199-208.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- HAUSMAN, J.A., and WISE, D.A. (1981). Stratification on endogenous variables and estimation; the Gary Income Maintenance Experiment. In *Structure Analysis of Discrete Data with Econometric Applications*, (Eds. C.F. Mansky and D. McFadden). Cambridge, Mass.: MIT Press, 366-391.
- HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A*, 143, 474-487.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B*, 42, 377-386.
- PATIL, G.P., and RAO, C.R. (1978). Weighted distributions and size biased sampling with application to wildlife populations and human families. *Biometrics*, 34, 179-189.
- PFEFFERMANN, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association*, 83, 824-833.
- PFEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review* (Forthcoming).
- RAO, C.R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions*, (Ed. G.P. Patil). Calcutta: Statistical Publishing Society, 320-332.
- RAO, C.R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? In *A Celebration in Statistics* (Eds. A.C. Atkinson and S.E. Fienberg). New York: Springer-Verlag, 543-569.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 469-474.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.
- RUBIN, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, (Eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith). Amsterdam: Elsevier Science, 463-472.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.
- SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10, 616-620.

Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used

CARL-ERIK SÄRNDAL¹

ABSTRACT

In almost all large surveys, some form of imputation is used. This paper develops a method for variance estimation when single (as opposed to multiple) imputation is used to create a completed data set. Imputation will never reproduce the true values (except in truly exceptional cases). The total error of the survey estimate is viewed in this paper as the sum of sampling error and imputation error. Consequently, an overall variance is derived as the sum of a sampling variance and an imputation variance. The principal theme is the estimation of these two components, using the data after imputation, that is, the actually observed values and the imputed values. The approach is model assisted in the sense that the model implied by the imputation method and the randomization distribution used for sample selection will together determine the appearance of the variance estimators. The theoretical findings are confirmed by a Monte Carlo simulation.

KEY WORDS: Single value imputation; Variance estimation; Imputation model; Model assisted inference.

1. DIFFERENT TYPES OF IMPUTATION

This paper reports work carried out in connection with the development of Statistics Canada's Generalized Estimation System (GES). Variance estimates are to be routinely calculated in the different estimation modules that define the GES. There was a need to develop suitable methods for variance estimation when the data set contains imputed values, which is the case in practically all surveys.

Two principal approaches to estimation with missing data are weighting and imputation. In the recent literature, the weights used to compensate for nonresponse are usually viewed as the inverse of the response probabilities associated with an assumed response mechanism. Since the response probabilities are ordinarily unknown, they need to be estimated from the available data. Imputation, on the other hand, has the advantage that it yields a complete data matrix. Such a matrix simplifies data handling, but it does not imply that "standard estimation methods" can be used directly. The imputed values are sample-based, thus they have their own statistical properties, such as a mean and a variance.

In our age, imputation is an extensively used tool. It is interesting to note what Pritzker, Ogus and Hansen (1965) say about imputation policy at the US Bureau of the Census: "Basically our philosophy in connection with the problem of . . . imputation is that we should get information by direct measurement on a very high proportion of the aggregates to be tabulated, with sufficient control on quality that almost any reasonable rule for . . . imputation will yield substantially the same results . . . With respect to imputation in censuses and sample surveys we have adopted a standard that says we have a low level of imputation, of the order of 1 or 2 percent, as a goal."

¹ Carl-Erik Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec) H3C 3J7.

Ideally, we should still strive for the goal of only one to two percent imputation. But in our time most surveys carried out by large survey organizations show a rate of imputation that is much higher. Clearly, if 30% of the values are imputed, the effects of imputation can not be ignored. Imputation can create systematic error (bias) in the point estimate; this is perhaps the most serious concern. But even if an imputation method can be found such that there is no appreciable systematic error, one must not ignore the often considerable effect that imputation has on the precision (the variance) of the point estimate. There is a need for simple yet valid variance estimation methods for survey data containing imputations, so that the coefficients of variation of the survey estimates can be properly reported.

A variety of imputation methods have been proposed. These can be classified in different ways. One way to classify is by the number of imputations carried out. In **single imputation** methods, a single value is imputed for a missing value. A complete data matrix is obtained, in which the imputed values are flagged. Estimates are calculated with the aid of the completed set. In **multiple imputation**, two or more values are imputed for each missing value. Several completed data sets are thus obtained. Estimates are calculated with the aid of the completed data sets.

Imputation methods also differ with respect to the modeling underlying the imputation. Some imputation methods use an **explicit** model, as when the imputed value is obtained by a regression fit, a ratio or mean imputation. In other methods, the model is only **implicit**, as in hot deck imputation and nearest neighbour donor imputation. The distinctions just made are important for this paper.

Statistics Canada currently uses imputation methods such as nearest neighbour donor, current ratio, current mean, previous value, previous mean, auxiliary trend. All of these are single imputation methods. The imputed values originate in the Generalized Edit and Imputation System (GEIS), from where they enter into the Generalized Estimation System (GES), where the point estimates and the variance estimates are calculated in a number of different estimation modules. This paper deals in particular with current ratio imputation, which represents a case of explicit modeling.

2. SOME THOUGHTS ON MULTIPLE IMPUTATION

Multiple imputation was suggested by D.B. Rubin around 1977. His ideas are explained in a number of papers, of which Herzog and Rubin (1983) and Rubin (1986) are expository, and in a book, Rubin (1987). Multiple imputation has advantages as well as disadvantages; the same is true for single imputation.

Rubin (1986) sees as a disadvantage of single imputation that "... the one imputed value cannot in itself represent uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reason for nonresponse are known."

Multiple imputation is attractive because it communicates the idea that imputation has variability. It is precisely this variability – the variability within and between the several completed data sets – that is exploited in the variance estimation methods proposed under multiple imputation. These methods make powerful use of basic statistical concepts. (On the other hand, one can argue that sample selection also has variability, but most surveys cannot afford more than a single sample, and estimation must be carried out with this unique sample.)

Simple examples show that treating imputed values just like observed values can lead to severe underestimation of the true uncertainty; survey samplers have long been aware of this. And

it is a fact that users sometimes treat imputed values just like observed values, with wrong statement of precision as a result. With modern computers, it is easy to impute by some rule or another, but not so easy to obtain valid variance estimates.

The citation above seems to conclude that because a single imputed value does not display variation, we cannot obtain reasonable variance estimates; we are necessarily led to underestimation. I do not share this opinion. The methods that I discuss show that valid variance estimation is indeed possible with single imputation.

A method for variance estimation in the presence of imputed values should have the following properties: (a) a sound theoretical backing; (b) robustness to the assumptions underlying the imputation; (c) it must be practical, easy to carry out, and readily accepted by users.

While multiple imputation has the ingredients (a) and (b), it is clear that, in some applications at least, it does not have the property (c). In the development of the GES we must depend on procedures that are easy to administer and easy to accept by the user. The user of a data set (someone who is not primarily a statistician) can easily understand that the statistician imputes once, with the objective to fill in the best possible value for one that is missing. While it is true that for some purposes, such as secondary analyses, it might be interesting to have several completed data matrices, the costs of storage of multiple data sets will often rule out this option.

Multiple imputation may well be useful in other contexts and for other reasons than those that are essential to the development of the GES. The multiple imputation method has indicated one way of handling the problem of understatement of the variance, at least for some situations. The method has recently come under criticism by Fay (1991) and is not the only answer. Let us see what can be done with single imputation methods. The method described below is based on Särndal (1990).

3. IMPUTATION VARIANCE AND SAMPLING VARIANCE

An imputation rule corresponds to an (explicit or implicit) model for the relationship among variables of interest to the survey. That is, when the analyst has fixed an imputation rule, he or she has in fact chosen a model. The principle for the developments that follow is that if this rule is considered good enough for the point estimates (no systematic error), the rule is also good enough for the corresponding estimates of variance. In other words, the model maker should take responsibility for control of the bias as well as for the appropriateness of the variance estimate.

Let $U = \{1, \dots, k, \dots, N\}$ be a finite population; let y denote one of the study variables in the survey. The objective is to estimate the population total of y , $t = \sum_U y_k$. (If C is any set of population units, where $C \subseteq U$, \sum_C is used as shorthand for $\sum_{k \in C}$, for example, $t = \sum_U y_k$ means $\sum_{k \in U} y_k$.) A probability sample s is selected with a given sampling design. The inclusion probabilities are known, and ordinary design-based variance estimates would be obtained if all units $k \in s$ are observed. However, there are missing data. Let r be the subset s for which the values y_k are actually observed. For the complement, $s - r$, imputations are calculated. The **data after imputation** consist of the values denoted $y_{\bullet k}$, $k \in s$, such that

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ y_{\text{imp},k} & \text{if } k \in s - r, \end{cases}$$

where y_k is an actually observed value, and $y_{\text{imp},k}$ denotes the imputed value for the unit k . The case $r = s$ implies no imputation; all data are actual observations.

Let us write the estimator of t that would be used in the case of 100% response (that is, $r = s$) as $\hat{t} = \sum_{k \in s} w_k y_k = \sum_s w_k y_k$, where w_k is the weight given to the observation y_k . For example, in simple random sampling without replacement (SRSWOR) of n units from N , $w_k = N/n$ for all $k \in s$ when the expanded sample mean is used to estimate t , and $w_k = (\bar{z}_U / \bar{z}_s)(N/n) = (\sum_{U \in k} z_k) / (\sum_s z_k)$ for all $k \in s$ when the ratio estimator is used with z as an auxiliary variable.

When the data contain imputations, the estimator of t is $\hat{t}_\bullet = \sum_s w_k y_{\bullet k}$. That is, we assume that the weights w_k are identical to those used when all data are actual observations. This principle is used in the estimation modules of the GES. It embodies an assumption that imputation by the chosen rule causes little or no systematic error in the estimates.

The variance of an estimated total is increased by imputation, because imputation does not (except in truly exceptional circumstances) reproduce the true value y_k . Concrete evidence of this is the fact that if the imputation rule is applied to the actually observed sample units, there will always be error. If the rule is not without error for the responding units, it is not without error for the nonresponding units either. In Section 4 we express the variance of \hat{t}_\bullet as a sum of two components, a sampling variance, and a variance due to imputation,

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}.$$

The imputation variance V_{imp} is zero if all data are actually observed values, or if the imputation procedure is capable of exactly reproducing the true value y_k for every unit requiring imputation. (Neither case is likely in practice.) The procedure given in Section 4 uses the data after imputation, $y_{\bullet k}$, $k \in s$, to obtain estimates of each of the two components, leading to

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}.$$

The component \hat{V}_{sam} is calculated in two steps:

- (1) Compute the standard design-based variance estimate using the data after imputation. (For example, if SRSWOR is used, and $r = s$, the standard unbiased variance estimate of $N\bar{y}_s$ is $N^2(1/n - 1/N) \sum_s (y_k - \bar{y}_s)^2 / (n - 1)$. This formula, calculated on the data after imputation, yields $N^2(1/n - 1/N) \sum_s (y_{\bullet k} - \bar{y}_{\bullet s})^2 / (n - 1)$, where $\bar{y}_{\bullet s}$ is the mean of the n values $y_{\bullet k}$.)
- (2) Add a term to correct for the fact that many imputation rules give data with “less than natural” variability, which would lead to understatement of the sampling variance unless corrective action is taken. Finally, the component \hat{V}_{imp} is readily computed from the data after imputation. The user will easily accept the argument that the variance obtained by the standard formula is not sufficient in itself; something must be added because the imputation rule is less than perfect.

The method has the good property that if no imputation is required, that is, $r = s$, then $\hat{V}_{\text{imp}} = 0$ and \hat{V}_{sam} equals the “standard variance estimator” that one would have used with 100% actually observed values.

4. THEORETICAL DEVELOPMENTS

The total error of \hat{t}_\bullet is decomposed as

$$\hat{t}_\bullet - t = (\hat{t} - t) + (\hat{t}_\bullet - \hat{t}) = \text{sampling error} + \text{imputation error}.$$

The imputation error is the difference between the unknown estimate that would have been calculated if the data had consisted entirely of actual observations and the estimate that can be calculated on the data after imputation. The imputation error is

$$\hat{t}_{\bullet} - \hat{t} = - \sum_{s=r} w_k e_k,$$

where

$$e_k = y_k - y_{\text{imp},k}$$

is an **imputation residual** which can not be observed for a unit $k \in s - r$. The magnitude of e_k depends on how well the imputation model fits. The residuals are small if the imputation method gives nearly perfect substitute values. To pursue the argument, different directions may be taken. Here, we use a **model assisted** approach in which three different probability distributions are considered. The corresponding expectation symbols are written as E_{ξ} , E_s , and E_r . Here, ξ indicates “with respect to the imputation model”; s indicates “with respect to the sampling design”, and r indicates “with respect to the response mechanism, given s ”. The model is implied by the imputation rule, so it is known; the sampling design is the given probability sampling distribution, so it is also known; the response mechanism is an ordinarily unknown distribution governing the response, given the sample s .

The estimator \hat{t}_{\bullet} is overall unbiased in the sense that $E_{\xi} E_s E_r (\hat{t}_{\bullet} - t) = 0$ if two conditions hold:

- the order of the expectation operators can be changed so that $E_{\xi} E_s E_r (\cdot)$ can be evaluated as $E_s E_r \{E_{\xi}(\cdot | s, r)\}$, and
- the imputation residual $e_k = y_k - y_{\text{imp},k}$ has zero model expectation for every $k \in r$, that is, $E_{\xi}(e_k) = 0$, which implies that $E_{\xi}(\hat{t}_{\bullet} - \hat{t}) = 0$.

Condition (a) is satisfied if the response mechanism is one that may depend on s and on auxiliary data, but not on the y -values, y_k , $k \in s$. That is, the probability $q(r)$ of realizing the response set r is of the form $q(r) = q(r | s, \{x_k: k \in s\})$, where $\{x_k: k \in s\}$ denote the auxiliary data. The response mechanism can then be said to be ignorable.

We now examine the overall variance given by

$$V_{\text{tot}} = E_{\xi} E_s E_r \{(\hat{t}_{\bullet} - t)^2\},$$

which may also be called the anticipated variance under the imputation model ξ . We obtain

$$\begin{aligned} V_{\text{tot}} &= E_{\xi sr}(\hat{t}_{\bullet}) = E_{\xi} E_s E_r \{(\hat{t}_{\bullet} - t)^2\} \\ &= E_{\xi} E_s E_r \{(\hat{t} - t) + (\hat{t}_{\bullet} - \hat{t})\}^2 \\ &= E_{\xi} V_p + E_s E_r V_{\xi c}, \end{aligned} \quad (4.1)$$

where $V_p = E_s \{(\hat{t} - t)^2\}$ is the design-based variance of \hat{t} , supposing \hat{t} is design unbiased for the total t . (For an estimator with some slight design bias, V_p is the design-based mean square error of \hat{t} .) Note that $(\hat{t} - t)$ depends on s only, and not on r . Moreover,

$$V_{\xi c} = E_{\xi} \{(\hat{t}_{\bullet} - \hat{t})^2 | s, r\}$$

is the model variance of the imputation error, conditionally on s and r . The subscript c stands for “conditional”. The derivation of (4.1) assumes that condition (a) holds so that the expectation E_ξ can be moved inside $E_s E_r$, and that the mixed term

$$2E_\xi E_s [(\hat{t} - t)\{E_r(\hat{t}_\bullet - \hat{t}) \mid s\}] \quad (4.2)$$

vanishes or is sufficiently close to zero that we can ignore it. This would be the case if the expected imputation error is zero or negligible under the response mechanism, conditionally on the realized probability sample s . Even if (4.2) is not exactly zero for the mechanism that determines the response, we can in many cases approximate (4.2) by zero and still use the method below to obtain a variance estimate that is much better than pretending naively that imputed data are as good as actually observed data. For ratio imputation and SRSWOR, which is an application considered in Section 5, the term (4.2) is exactly zero.

If we denote $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_s E_r V_{\xi c}$ in (4.1), then

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$$

or

$$\text{overall variance} = \text{sampling variance} + \text{imputation variance}.$$

The objective is to estimate the overall variance, so that a valid confidence interval for the unknown t can be calculated. Our approach is to obtain separate estimates, \hat{V}_{sam} and \hat{V}_{imp} , of the two components $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_s E_r V_{\xi c}$. The data available for this estimation are $y_{\bullet k}$, $k \in s$. The argument for obtaining \hat{V}_{sam} and \hat{V}_{imp} is as follows:

- (i) Estimation of the sampling variance component. Let \hat{V}_p be the standard (design-unbiased or nearly design-unbiased) estimator of the design variance V_s . Denote by $\hat{V}_{\bullet p}$ the quantity obtained by calculating \hat{V}_p from the data after imputation, $y_{\bullet k}$, $k \in s$. For many imputation rules, $\hat{V}_{\bullet p}$ underestimates V_{sam} . The underestimation is compensated in the following way. Evaluate the conditional expectation

$$E_\xi(\hat{V}_p - \hat{V}_{\bullet p} \mid s, r) = V_{\text{dif}}.$$

Then for given s and r , find a model unbiased estimator, denoted \hat{V}_{dif} , of V_{dif} . This will usually require the estimation of certain parameters of the model ξ . Consequently,

$$E_\xi(\hat{V}_{\text{dif}} \mid s, r) = E_\xi(\hat{V}_p - \hat{V}_{\bullet p} \mid s, r).$$

Then

$$\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}}$$

is overall unbiased for the component $V_{\text{sam}} = E_\xi V_p$, as the following derivation shows:

$$\begin{aligned} E_\xi E_s E_r(\hat{V}_{\text{sam}}) &= E_s E_r \{E_\xi(\hat{V}_{\bullet p}) + E_\xi(\hat{V}_{\text{dif}})\} \\ &= E_s E_r \{E_\xi(\hat{V}_p)\} = E_\xi E_s(\hat{V}_p) \\ &= E_\xi V_p = V_{\text{sam}}. \end{aligned}$$

- (ii) Estimation of the imputation variance component. Simply find an estimator, $\hat{V}_{\xi c}$, that is model unbiased for $V_{\xi c}$. That is, $E_{\xi}(\hat{V}_{\xi c}) = V_{\xi c}$. Again, this may require the estimation of unknown parameters of the model ξ . Then $\hat{V}_{\xi c}$ is overall unbiased for the imputation variance component V_{imp} , since

$$E_s E_r E_{\xi}(\hat{V}_{\xi c}) = E_s E_r V_{\xi c} = V_{\text{imp}}.$$

Finally, an overall unbiased estimator of V_{tot} is given by

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}},$$

where $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}}$ and $\hat{V}_{\text{imp}} = \hat{V}_{\xi c}$. Note that the role of \hat{V}_{dif} is to correct for the fact that the data after imputation may display “less than natural” variation. This often happens when $y_{\text{imp},k}$ equals the predicted value from a fitted regression, that is, “the value on the line”. The variation around the line is not reflected in the predicted value.

To be overall unbiased, the estimator \hat{V}_{tot} constructed above requires that condition (a) holds, that (4.2) is zero, and that the imputation model is correct, so that \hat{V}_{dif} and $\hat{V}_{\xi c}$ are model unbiased for V_{dif} and $V_{\xi c}$, respectively. Mild departures from the assumed imputation model may not have serious consequences, but if the imputation model is grossly misspecified it is clear that \hat{V}_{tot} may be considerably biased because of the model bias of \hat{V}_{dif} and $\hat{V}_{\xi c}$. Monte Carlo simulations reported in Lee, Rancourt and Särndal (1992) show that the variance estimator \hat{V}_{tot} is fairly robust to imputation model breakdown. To add the terms \hat{V}_{dif} and $\hat{V}_{\xi c}$ is in any case a vast improvement on simply using the naive uncorrected variance estimator $\hat{V}_{\bullet p}$.

Note that if the imputation model holds, an unbiased variance estimate is obtained with the method even if the response probabilities differ among units, as long as they depend on the x_k -values only. That is, we can allow a systematic response pattern such that large x_k -value units are less likely to respond than small x_k -value units. If the response probabilities depend explicitly the y_k -values, then the situation is different; the response mechanism is nonignorable and condition (a) does not hold. There will now be bias in \hat{V}_{tot} due to nonignorability; the simulations in Lee, Rancourt and Särndal (1992) throw some light on the magnitude of this bias.

Example. The sample s is drawn with SRSWOR; n units from N . Let m denote the size of the response set r . Suppose the respondent mean is imputed for units requiring imputation. The corresponding imputation model ξ states that $y_k = \beta + \epsilon_k$, where the ϵ_k are uncorrelated errors terms with $E_{\xi}(\epsilon_k) = 0$, $V_{\xi}(\epsilon_k) = \sigma^2$. That is, $y_{\bullet k} = y_k$ if $k \in r$ and $y_{\bullet k} = \hat{\beta} = \bar{y}_r$ if $k \in s - r$, and we obtain the estimator $\hat{t}_{\bullet} = (N/n) \sum_s y_{\bullet k} = N\bar{y}_r$. Here the standard design-based variance estimator for 100% response is $\hat{V}_p = N^2(1/n - 1/N) \sum_s (y_k - \bar{y}_s)^2 / (n - 1)$; when this formula is computed on data after imputation we get $\hat{V}_{\bullet p} = N^2(1/n - 1/N) \{ (m - 1)/(n - 1) \} S_{yr}^2$, where $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2 / (m - 1)$. Other derivations give $\hat{V}_{\text{dif}} = N^2(1/n - 1/N) \{ (n - m)/(n - 1) \} S_{yr}^2$ and $\hat{V}_{\text{imp}} = N^2(1/m - 1/n) S_{yr}^2$. Thus, $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}} = N^2(1/n - 1/N) S_{yr}^2$, and $\hat{V}_{\text{tot}} = N^2(1/m - 1/N) S_{yr}^2$, which is easy to accept as a “good” variance estimator for this simple imputation rule. The following table shows the contribution of each of the three terms to the total variance estimator \hat{V}_{tot} , for different rates of imputation, assuming that N is large compared to m and n , and $(m - 1)/m \approx (n - 1)/n \approx 1$.

Imputation rate in %	% contribution to \hat{V}_{tot}		
	$\hat{V}_{\bullet,p}$	\hat{V}_{dif}	\hat{V}_{imp}
100 (1 - m/n)			
10	81	9	10
20	64	16	20
30	49	21	30

The table illustrates the dangers of acting as if imputations are real data: with 30% imputed values, the standard formula variance estimator $\hat{V}_{\bullet,p}$ in this example covers less than half of the correctly estimated total variance. Imputation by the respondent mean is useful as an example; the results are particularly simple. But usually in practice, respondent mean imputation is neither justified nor efficient. The underlying model is not sophisticated enough to avoid systematic error in the point estimates, and the residuals $e_k = y_k - \bar{y}_r$ can vary considerably.

5. APPLICATION TO IMPUTATION BY THE CURRENT RATIO METHOD

The method assumes that a positive auxiliary value x_k is known for every unit $k \in s$. If $k \in s - r$, we impute $y_{\text{imp},k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k) / (\sum_r x_k)$. The data after imputation are

$$y_{\bullet,k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{B}x_k & \text{if } k \in s - r. \end{cases}$$

The model behind current ratio imputation is

$$y_k = \beta x_k + \epsilon_k, \tag{5.1}$$

where the ϵ_k are uncorrelated model errors such that

$$E_{\xi}(\epsilon_k) = 0, \quad V_{\xi}(\epsilon_k) = \sigma^2 x_k. \tag{5.2}$$

Suppose that the sample s is selected by SRSWOR. Let the respective sizes of s , r , and $s - r$ be n , m , and $n - m$. If no imputation was needed, the estimator of $t = \sum_U y_k$ would be $\hat{t} = N\bar{y}_s$. Using the data after imputation, we get

$$\hat{t}_{\bullet} = (N/n) \sum_s y_{\bullet,k} = N\bar{x}_s \bar{y}_r / \bar{x}_r. \tag{5.3}$$

(Overbar and subscript s , r , or $s - r$ indicates “straight mean”, for example, $\bar{y}_r = \sum_r y_k / m$, $\bar{x}_{s-r} = \sum_{s-r} x_k / (n - m)$, etc.) Using the results of the preceding section, we have $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$ with $V_{\text{sam}} = E_{\xi}\{N^2(1/n - 1/N)S_{yU}^2\}$ and $V_{\text{imp}} = E_s E_r\{N^2(1/m - 1/n)C_1\sigma^2\}$, where $S_{yU}^2 = \sum_U (y_k - \bar{y}_U)^2 / (N - 1)$ and $C_1 = \bar{x}_s \bar{x}_{s-r} / \bar{x}_r$, a known constant. The mixed term (4.2) is exactly zero in this case. Our method of variance estimation gives $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$, where

$$\hat{V}_{\text{sam}} = N^2(1/n - 1/N)\{S_{y_{\bullet}s}^2 + C_0\hat{\sigma}^2\}, \tag{5.4}$$

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)C_1\hat{\sigma}^2, \quad (5.5)$$

where $S_{y_{\bullet s}}^2 = \sum_s (y_{\bullet k} - \bar{y}_{\bullet s})^2 / (n - 1)$ is the variance calculated on data after imputation, and we have chosen to estimate σ^2 by the model unbiased formula

$$\sigma^2 = \frac{1}{\bar{x}_r \{1 - (1/m)(cv_{xr})^2\}} \frac{\sum_r (y_k - \hat{B}x_k)^2}{m - 1},$$

where $cv_{xr} = S_{xr}/\bar{x}_r$ is the coefficient of variation of x in the response set r . The constant C_0 is obtained as

$$C_0 = \frac{1}{\sigma^2} E_{\xi} (S_{ys}^2 - S_{y_{\bullet s}}^2),$$

where

$$S_{ys}^2 = \frac{1}{n - 1} \sum_s (y_k - \bar{y}_s)^2$$

is the (unknown) sample variance based on data with 100% actual observations. After evaluation,

$$C_0 = \frac{1}{n - 1} \left\{ \sum_{s-r} x_k - \frac{\sum_{s-r} x_k^2}{\sum_r x_k} + \frac{1}{n} \frac{\sum_{s-r} x_k \sum_s x_k}{\sum_r x_k} \right\}.$$

If m is not too small, the approximations $\hat{\sigma}^2 \approx (\sum_r e_k^2) / (\sum_r x_k)$ with $e_k = y_k - \hat{B}x_k$ and $C_0 \approx (1 - m/n)\bar{x}_{s-r}$ are sufficiently good for most applications.

We can write the imputation variance component as

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)A\bar{x}_s\hat{\sigma}^2,$$

where $A = \bar{x}_{s-r}/\bar{x}_r$. The constant A reflects the selection effect due to nonresponse. If large units are less inclined to respond than small units, then A may be considerably greater than unity, and, for a given a sample s and a given number m of respondents, the component \hat{V}_{imp} tends to be large, relative to a case where, say, all units are equally likely to respond. This tendency makes good sense intuitively.

Two special cases are noted: (1) If all $x_k = 1$, the estimated total variance becomes simply

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}} = N^2(1/m - 1/N)S_{yr}^2,$$

where S_{yr}^2 is the variance of the m actual observations y_k . This agrees with the variance obtained under a two-phase sampling design with SRSWOR in each phase. (2) If no imputation is required, that is, if $s = r$, then $\hat{V}_{\text{imp}} = 0$, and

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} = N^2(1/n - 1/N)S_{ys}^2.$$

That is, our method yields the well known variance estimator for SRSWOR.

A Monte Carlo study with 100,000 repeated response sets r was carried out to confirm the above results for current ratio imputation. A finite population of size $N = 100$ was generated according to the model consisting of (5.1) and (5.2). The typical response set r was obtained

as follows: Draw a SRSWOR sample s of size $n = 30$; given s , generate r by a response mechanism in the form of independent Bernoulli trials, one for each $k \in s$, with probability θ_k for the outcome “response”. Three different response mechanisms were used: Mechanism 1: θ_k increases with y_k in such a way that $\theta_k = 1 - \exp(-a_1 y_k)$; Mechanism 2: θ_k increases as y_k decreases in such a way that $\theta_k = \exp(-a_2 y_k)$; Mechanism 3: θ_k is constant at 0.7, that is, a uniform response mechanism. The constants a_1 and a_2 in the first two response mechanisms (which can be described as non-ignorable) were fixed to obtain an average response probability of 0.7. The sizes of the realized response sets r thus varied around a mean of 21 for all three mechanisms. For each r , the point estimate \hat{t}_\bullet , given by (5.3) was calculated as well as three different variance estimators, $\hat{V} = \hat{V}(\hat{t}_\bullet)$. These were: (1) the **model assisted** variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ equal to the total of (5.4) and (5.5); (2) the **two-phase** sampling variance estimator $N^2(1/n - 1/N)S_{y_r}^2 + N^2(1/m - 1/n)\sum_r e_k^2/(m - 1)$, an estimator which follows from standard two-phase sampling theory with an assumption of SRSWOR subsampling of m respondents from the n units in the initial sample (Rao 1990); and (3) the **standard unadjusted** variance estimator $N^2(1/n - 1/N)S_{y_\bullet s}^2$ obtained by acting as if imputations are as good as actual data. The results are shown in the following table.

Estimator \hat{V}	Relative bias of \hat{V} in %		
	Mechanism 1	Mechanism 2	Mechanism 3
Model assisted	-0.20	-4.64	-3.99
Two-phase	9.95	-12.49	-1.11
Standard unadjusted	-25.73	-37.90	-33.21

The relative bias of an estimator \hat{V} was calculated as $\{\text{mean}(\hat{V}) - \text{var}(\hat{t}_\bullet)\}/\text{var}(\hat{t}_\bullet)$, where $\text{mean}(\hat{V})$ is the mean of the 100,000 values of \hat{V} , and $\text{var}(\hat{t}_\bullet)$ is the variance of the 100,000 values of \hat{t}_\bullet . The simulation shows that the model assisted variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ is nearly unbiased for all three response mechanisms. In a way, this is not surprising because the population was generated to agree with the ratio imputation model. Mechanisms 1 and 2 are of the nonignorable kind and do not verify condition (a) of Section 4 required for unbiasedness of \hat{V}_{tot} . Interestingly, though, in this example the bias of \hat{V}_{tot} remains small despite this. The two-phase estimator works well for the uniform response mechanism 3, the case for which it was conceived; otherwise it is biased. Finally, to act as if imputed data are as good as actual data leads, as expected, to a dramatic understatement of the true variance for all three mechanisms. A more extensive Monte Carlo study of ratio estimation is reported in Lee, Rancourt and Särndal (1992). This paper gives an idea of the effect of imputation model misspecification, which is also discussed in Rao (1992).

6. IMPUTED VALUES THAT HAVE AN ADDED RESIDUAL

We can distinguish two types of imputed values: (1) the imputed value $y_{\text{imp},k}$ consists of a predicted value only, $y_{\text{pred},k}$, as when the value on a fitted regression line or surface is used. For example in the current ratio imputation method as used above, $y_{\text{imp},k} = y_{\text{pred},k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$; (2) the imputed value $y_{\text{imp},k}$ consists of a predicted value and a

residual, so that $y_{\text{imp},k} = y_{\text{pred},k} + e_k^*$. The residual term, whose purpose is to make imputed values more like actual observations, may be obtained by sampling the residuals $e_k = y_k - y_{\text{pred},k}$ calculated for the responding units $k \in r$. A scheme for this is given below. This type of imputation is sometimes recommended in the literature as a means of preserving the distributions of the imputed data; see, for example, the discussion in Little (1988). The imputation process then requires more effort to complete, and for the purposes of the GES (whose principal aim is valid estimation of the precision of survey estimates), it is not clear that the advantages gained are worth the extra effort.

Let us, however, indicate one scheme for imputation by "predicted value plus residual" in the case where the current ratio imputation model is taken as the point of departure: For $k \in r$, calculate $e_k = y_k - \hat{B}x_k$ with $\hat{B} = (\sum_r y_k) / (\sum_r x_k)$, then $\tilde{e}_k = e_k / \sqrt{x_k}$. This gives a supply of m "standardized residuals" \tilde{e}_k . Then for a unit $k \in s-r$, calculate $e_k^0 = \sqrt{x_k} \tilde{e}_k$, where \tilde{e}_k is drawn by SRSWR from the supply, and x_k belongs to the unit requiring imputation. Then large x -value units tend to obtain larger residuals e_k^0 , which is consistent with the model. Then set $e_k^* = e_k^0 - (\sum_{s-r} e_k^0) / (n - m)$. For $k \in s-r$, impute $y_{\text{imp},k} = \hat{B}x_k + e_k^*$, $k \in s-r$; for $k \in r$, we have actual observations, y_k . Since the e_k^* were made to sum to zero over $s - r$, the point estimator is given by $\hat{t}_\bullet = (N/n) \sum_s y_{\bullet k} = N \bar{x}_s \bar{y}_r / \bar{x}_r$ as in Section 5, but its variance is different. It can be shown that $E_\xi E_s E_r E_\# (S_{y_\bullet s}^2 - S_{y_s}^2) \approx 0$, where $E_\#$ denotes average with respect to the random selection of a standardized residual. That is, the difference between the variance calculated on data after imputation, $S_{y_\bullet s}^2$, and the unknown variance of a sample consisting entirely of actual observations, $S_{y_s}^2$, is approximately zero on the average. We can use $\hat{V}_{\text{sam}} = N^2(1/n - 1/N)S_{y_\bullet s}^2$ as an approximately overall unbiased estimator of the sampling variance component. There is no need now to add a correction \hat{V}_{dif} . However, an estimator of the imputation variance $V_{\text{imp}} = N^2(1/m - 1/n)C_1 \sigma^2$ must still be calculated and added to \hat{V}_{sam} .

7. CONCLUDING REMARKS

The continued work on the variance estimation techniques outlined in this paper has the following objectives: (1) extensions to imputation procedures based on models that are implicit only, in particular the nearest neighbour donor method; (2) extensions to the case where there is a mixture of several imputation procedures in the same survey.

Deville and Särndal (1992) present results for an extension in which the Horwitz-Thompson estimator, $\hat{t} = \sum_s y_k / \pi_k$, serves as the prototype. The estimator using data after imputation is then

$$\hat{t}_\bullet = \sum_r y_k / \pi_k + \left(\sum_{s-r} x_k / \pi_k \right)' \hat{B} = \sum_s y_k / \pi_k - \sum_{s-r} e_k / \pi_k,$$

where $e_k = y_k - x_k' B$ is the imputation residual for unit k obtained by multiple regression.

ACKNOWLEDGEMENTS

I am indebted to M. Hidirolou, P. Lavallée, Y. Leblond, H. Lee, and G. Reinhardt of Statistics Canada for their collaboration in the work that led to this paper. The comments of two referees led to improvements in the original manuscript and are gratefully acknowledged.

REFERENCES

- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Variance estimation for survey data with regression imputation. Technical report.
- HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in surveys. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 209-245.
- FAY, R.E. (1991). A design-based perspective on missing data variance. Proceedings, 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.
- LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1992). Experiments with variance estimation from survey data with imputed values. Report, Business Survey Methods Division, Statistics Canada, submitted for publication.
- LITTLE, R.J.A. (1988). Missing-data adjustments in large surveys (with discussion). *Journal of Business and Economic Statistics*, 6, 287-301.
- PRITZKER, L., OGUS, J., and HANSEN, M.H. (1965). Computer editing methods: some applications and results. *Bulletin of the International Statistical Institute*, 41, 442-466.
- RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Manuscript seen by courtesy of the author.
- RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Manuscript seen by courtesy of the author.
- RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- SÄRNDAL, C.-E. (1990). Estimation of precision in the generalized estimation system when imputation is used. Report, Informatics and Methodology Field, Statistics Canada, March 31, 1990.

A Sample Allocation Method for Two-Phase Survey Designs

J.B. ARMSTRONG and C.F.J. WU¹

ABSTRACT

Motivated by a business survey design at Statistics Canada, we formulate the problem of sample allocation for a general two-phase survey design as a constrained nonlinear programming problem. By exploiting its mathematical structure, we propose a solution method that consists of iterations between two subproblems that are computationally much simpler. Using an approximate solution as a starting value, the proposed method works very well in an empirical study.

KEY WORDS: Optimal allocation; Convex programming.

1. INTRODUCTION

The purpose of this paper is to propose a method of sample allocation for two-phase survey designs. Suppose it is necessary to stratify a population of size N into L strata according to an auxiliary variable, z , whose information is not known before sampling. Values of a second auxiliary (size) variable, x , that is correlated with the variable of interest, y , are known for all units in the population. At the first phase of sampling, the population is divided into G strata according to x . An initial sample is drawn from size stratum g ($g = 1, 2, \dots, G$), using simple random sampling with sampling fraction v_g , and the z -value for each sampled unit is observed. At the second phase, units in the sample from size stratum g with z -value in class h ($h = 1, 2, \dots, L$), are subsampled using sampling fraction v_{gh} . The value of y is observed for units in the second-phase sample.

In the case of no size stratification ($G = 1$) Cochran (1977) gives the allocation that minimizes the variance of the estimate $\hat{Y} = \sum_h \sum_{i \in s2 \cap h} y_i / (v \cdot v_h)$ of the population total $Y = \sum_h N_h \cdot \bar{Y}_h$, subject to a fixed survey cost, C , where N_h and \bar{Y}_h are the population size and population mean, respectively, for stratum h and $\sum_{i \in s2 \cap h} y_i$ denotes the sum of y -values for units in the second phase sample, $s2$, with z -value in class h . If survey estimates are used for analytical purposes, the variance of the estimated total for z class h , $\hat{Y}_h = \sum_{i \in s2 \cap h} y_i / (v \cdot v_h)$, is also of interest. Sedransk (1965), Booth and Sedransk (1969), Rao (1973) and Smith (1989) have studied allocation problems involving the minimization of a function of variances of estimated class totals, subject to a cost constraint.

The method described in this paper can be used to solve the allocation problem for general G when there is a constraint on the variance of the estimated total for each z class. The method was motivated by an application in a business survey conducted by Statistics Canada. The survey involves the sampling of tax records for businesses.

Information about the population of taxfilers is made available to Statistics Canada by Revenue Canada. There is a requirement to produce estimates of financial variables for domains defined by a cross-classification of four-digit Standard Industrial Classification (SIC4) and province. Only two digits of SIC are coded by Revenue Canada with sufficient accuracy. In

¹ J.B. Armstrong, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6 and C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1.

order to standardize the precision of estimates for SIC4 domains within each province, a two-phase sample design was implemented. The first-phase sample of taxfilers is selected at Revenue Canada using strata defined using SIC2 and gross business income (size). Before the second phase sample is selected, an SIC4 code, considered more accurate than codes available from Revenue Canada, is assigned to each sampled unit by Statistics Canada. Strata defined using SIC4 and size are employed during selection of the second-phase sample. The same size boundaries are used for both phases of sampling. A detailed description of the sample design can be found in Choudhry, Lavallée and Hidiroglou (1989b).

First-phase sample selection is done using Bernoulli sampling (also called Poisson sampling). Suppose that taxfiler i falls in first-phase stratum g within a particular province \times SIC2 cell. To determine whether taxfiler i is included in the first-phase sample, a pseudo-random number in the interval $(0,1)$, say R_i , is generated using the taxfiler's unique identification number. The taxfiler is included in the first-phase sample if $R_i \in (0, v_g)$. Bernoulli sampling based on a different set of pseudo-random numbers is used to select the second-phase sample. Using Bernoulli sampling, selection and processing can begin before complete information about the taxfiler universe is available. This advantage of Bernoulli sampling is important, since taxfiler universe information is accumulated over a two-year period. Sample sizes obtained using Bernoulli sampling are random. Choudhry, Lavallée and Hidiroglou (1989b) derive the variance of $\hat{Y}_{h-STRAT} = \sum_g \sum_{i \in s2 \cap g \cap h} y_i / (v_g \cdot v_{gh})$ using simple random sampling as an approximation to Bernoulli sampling as discussed in Sunter (1986). Under the approximation, a simple random sample of fixed size $n'_g = v_g \cdot N_g$ is selected in size stratum g at the first phase. Let n'_{gh} denote the number of units with SIC4 h in the first-phase sample for size stratum g . At the second phase, a simple random sample of size $n_{gh} = v_{gh} \cdot n'_{gh}$ is selected for SIC4 h and size stratum g , with v_{gh} considered fixed. The variance of $\hat{Y}_{h-STRAT}$ is given by

$$V_h = \sum_g \left(\frac{1}{v_g \cdot v_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{v_g} - 1 \right) \cdot B_{gh},$$

where

$$A_{gh} = N_{gh} \cdot S_{gh}^2,$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y_{gh}^2}{N_{gh}} - S_{gh}^2 \right),$$

and S_{gh}^2 is the population variance in the second-phase SIC4 \times size stratum gh .

The plan of the paper is as follows. In Section 2, the optimal allocation problem is formulated in the context of the two-phase tax sample. An iterative solution procedure, called the exact method, is proposed. Section 3 includes a description of an approximation to the optimal allocation that can be used to obtain starting values for the exact method. The results of an empirical study involving comparison of various starting values for the exact method are reported in Section 4. Section 5 concludes the paper.

2. EXACT METHOD

In this section the optimal allocation problem is described and an iterative solution method, called the exact method, is proposed. To formulate the problem in the context of two-phase tax sampling, it is sufficient to consider one SIC2 cell in a particular province containing N

units. The cost of selecting a unit in the first-phase sample is K_1 , regardless of the stratum in which the unit falls, while the cost of selecting a unit in the second-phase sample is K_2 , regardless of stratum. Under Bernoulli sampling, the cost function is

$$F^* = K_1 \cdot \sum_g n'_g + K_2 \cdot \sum_g \sum_h n_{gh}.$$

Since sample sizes n'_g and n_{gh} are random, we use the expected cost

$$F = K_1 \cdot \sum_g v_g \cdot N_g + K_2 \cdot \sum_g \sum_h v_g \cdot v_{gh} \cdot N_{gh}. \quad (1)$$

Rao (1973) and Smith (1989) also solve allocation problems for two-phase sample designs using expected values of random cost functions. In the tax sampling context, the total cost for a province is the sum of the costs for all SIC2 cells within the province. The estimated coefficient of variation of the cost of two-phase tax sampling for the province of Quebec, calculated using 1988 data, was about 1.85%. Coefficients of variation for overall (national) costs were smaller.

It is necessary to minimize (1) with respect to v_g , $g = 1, 2, \dots, G$, and v_{gh} , $g = 1, 2, \dots, G$, $h = 1, 2, \dots, H$ under the constraints

$$\sum_g \left(\frac{1}{v_g \cdot v_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{v_g} - 1 \right) \cdot B_{gh} \leq C_h^2 \cdot Y_h^2, \quad h = 1, 2, \dots, H, \quad (2)$$

$$0 < v_g \leq 1, \quad g = 1, 2, \dots, G,$$

$$0 < v_{gh} \leq 1, \quad g = 1, 2, \dots, G, \quad h = 1, 2, \dots, H,$$

where C_h denotes the target coefficient of variation for SIC4 domain h .

Attempts at direct solution of this problem using the IMSL (1987) implementation of the successive quadratic programming algorithm of Schittkowski (1985) produced mixed results. The algorithm worked well for problems with small numbers of variables and constraints. However, satisfactory solutions for problems including more than approximately 35 variables or more than approximately 50 constraints could not be obtained.

Some costs obtained using direct application of Schittkowski's algorithm in the tax sampling context are given in Table 1. The algorithm was applied to the allocation problems for some SIC2 cells in the province of Quebec involving large numbers of variables and/or constraints using data for tax year 1988. All first-phase and second-phase sampling fractions were started at one when the direct approach was used. The lowest cost obtained using the method that we call the exact method, which will be described later in this section, is also given. The information in the table indicates that direct use of the IMSL implementation of Schittkowski's algorithm is an inappropriate strategy for SIC2 cells with large numbers of variables and constraints.

The exact method is based on a substantial simplification of the problem defined by (1) and (2) that can be achieved by exploiting its structure. In particular, we divide the problem into two main steps that can be solved iteratively. At the first step, (1) is minimized with respect to v_g , $g = 1, 2, \dots, G$, conditional on values for all second-phase sampling fractions. This

Table 1
Results for Direct and Exact Methods

SIC 2	No. of variables	No. of constraints	Cost (\$) – direct	Cost (\$) – exact
30	62	86	5155**	1897
35	37	51	551	512
39	38	50	1667	1450
427*	39	48	27528**	3383

* Three digits of SIC are used for first-phase stratification for construction industries.
** The IMSL routine terminated with an internal error that could not be rectified after consulting published documentation.

step requires the use of nonlinear optimization techniques. The second step involves minimizing (1) with respect to the second-phase sampling fractions, conditional on the values of the first-phase sampling fractions obtained in the first step. No iterations are required for this minimization, since it has a closed form solution. Furthermore, it can be done independently for each $h = 1, 2, \dots, H$. After completion of the second step, the first step is repeated and the iterative process continued. Convergence is declared when changes in the cost function between consecutive iterations are small.

Let $v_g^{(i)}$ and $v_{gh}^{(i)}$ denote the estimates of the optimal values of v_g and v_{gh} obtained after i iterations (each iteration including one repetition of the two steps described above). At the beginning of iteration $i + 1$, the transformation of variables given by $X_g^{(i+1)} = 1/v_g^{(i+1)} - 1$ is required. This transformation redefines the optimization problem involved in the first step of the iteration as a problem with linear constraints and a convex objective function. Such a convex programming problem is easier to solve.

More precisely, each iteration involves:
(i) Minimization of

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h v_{gh}^{(i-1)} \cdot N_{gh} \right) / (X_g^{(i)} + 1)$$

with respect to $X_g^{(i)}$, $g = 1, 2, \dots, G$, subject to the constraints

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(i)} + 1}{v_{gh}^{(i-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(i)} \cdot B_{gh} \geq 0, \quad h = 1, 2, \dots, H$$
$$X_g^{(i)} \geq 0, \quad g = 1, 2, \dots, G.$$

(ii) Calculation of $v_g^{(i)} = 1/(X_g^{(i)} + 1)$, $g = 1, 2, \dots, G$. Minimization, independently for each $h = 1, 2, \dots, H$, of

$$F_h = \sum_g v_g^{(i)} \cdot v_{gh}^{(i)} \cdot N_{gh}$$

with respect to $v_{gh}^{(i)}$, $g = 1, 2, \dots, G$, subject to the constraints

$$C_h^2 \cdot \hat{Y}_h^2 - \sum_g \left(\frac{1}{v_g^{(i)} \cdot v_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^{(i)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < v_{gh}^{(i)} \leq 1, \quad g = 1, 2, \dots, G,$$

where h is considered fixed.

It will be shown in Section 3 that solution of step (ii) does not require use of numerical methods. Therefore, the exact method only requires the solution of a series of convex programming problems, each involving only G variables. A convex programming problem is much easier to solve than a general nonlinear programming problem. A local solution of a convex programming problem is also a global solution.

Let $F^{(i)}$ denote the value of the cost function, (1), obtained using $v_g^{(i)}$ and $v_{gh}^{(i)}$. The $F^{(i)}$ values form a monotonically decreasing sequence and therefore converge to a limit. Whether this limit value and the corresponding sampling fractions give the global minimum depends on the starting value. This problem is caused by the geometry of the constraints in (2). In practice one should try several starting values to get the best solution. One starting value is given by the approximate method, which is described in the next section and does not require iterations.

3. APPROXIMATE METHOD

In this section, an allocation method that gives an approximation to the optimal allocation is described. The method was first suggested by Choudhry, Lavallée and Hidirolou (1989a). Assuming that all the second-phase sampling fractions are equal to one, an approximation to the optimal allocation of the first-phase sample is calculated. Then the second-phase sample is allocated, conditional on the first-phase sampling fractions. Since the cost of sampling a unit in both phases of sampling does not depend on the stratum in which the unit falls, minimizing cost is equivalent to minimizing sample size at each step of this method.

At the first step of the method, an approximate solution to the optimal allocation problem for a one-phase sample design is calculated. This step involves finding the minimum, independently for each h , of

$$F^{(h)} = \sum_g v_{g|h} \cdot N_g \quad (3)$$

with respect to $v_{g|h}$, $g = 1, 2, \dots, G$. The notation $v_{g|h}$ is used to denote the fact that a sampling fraction for size stratum g is determined subject to only one precision constraint, namely the constraint for SIC4 domain h , where h is fixed. In particular, the minimization must be done subject to the constraints

$$\sum_g \left(\frac{1}{v_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \quad (4)$$

$$0 < v_{g|h} \leq 1, \quad g = 1, 2, \dots, G. \quad (5)$$

One can show that the minimum of (3) is obtained when (4) holds with equality, so that the problem defined by (3), (4), and (5) is equivalent to finding the critical point of the lagrangian

$$L = \sum_g v_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \right].$$

Setting the derivatives with respect to $v_{g|h}$ equal to zero yields

$$v_{g|h} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot (-\lambda)^{1/2}, \quad g = 1, 2, \dots, G. \quad (6)$$

Setting $\partial L / \partial \lambda = 0$ we obtain

$$(-\lambda)^{1/2} = \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} / \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \quad (7)$$

After substitution of (7) into (6), we obtain the optimal sampling fraction for size stratum g given only one precision constraint, for SIC4 domain h ,

$$v_{g|h}^* = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} / \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \quad (8)$$

If one or more of the sampling fractions given by (8) are greater than one, one can set them equal to one and solve a modified allocation problem with a reduced number of strata. This approach corresponds to the overallocation procedure discussed by Cochran (1977). It is necessary to calculate (8) for $h = 1, 2, \dots, H$. The approximate first-phase sampling fraction for size stratum g , v_g^* , is set equal to the largest value in the set $\{v_{g|h}^*, h = 1, 2, \dots, H\}$ for $g = 1, 2, \dots, G$, an approach that ensures that the precision constraint for each SIC4 domain will be satisfied.

Given first-phase sampling fractions, optimal second-phase sampling fractions can be easily determined. Assume that, for the SIC2 \times province cell h , the size strata included in the allocation problem correspond to a set of integers, Γ . We set the second-phase sampling fractions equal to one for those size strata that are not included in the allocation problem. Normally, one would have $\Gamma = \{1, 2, \dots, G\}$ but because of overallocation during allocation of the second-phase sample, for example, Γ may not include all integers between 1 and G . The problem of allocating the second-phase sample is equivalent to the problem of finding the minimum of

$$F_h = \sum_{g \in \Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} \quad (9)$$

with respect to v_{gh} , $g \in \Gamma$, subject to the constraints

$$\sum_{g \in \Gamma} \left(\frac{1}{v_{gh}} - 1 \right) \cdot \frac{A_{gh}}{v_g^*} \leq M_h, \quad (10)$$

$$0 < v_{gh} \leq 1, \quad g \in \Gamma, \quad (11)$$

where

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_g^*} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Note that the expected number of units with SIC4 h in the second-phase sample for size stratum g , $v_g^* \cdot N_{gh}$, is employed in (9). It is easy to show that (9) attains a minimum when the constraint (10) holds with equality. Consequently, the minimization problem is equivalent to finding the critical point of the lagrangian

$$L_h = \sum_{g \in \Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} + \lambda \cdot \left(M_h - \sum_{g \in \Gamma} \left(\frac{1}{v_{gh}} - 1 \right) \cdot \frac{A_{gh}}{v_g^*} \right),$$

with respect to and v_{gh} , $g \in \Gamma$, and λ , subject to the constraints

$$0 < v_{gh} \leq 1, \quad g \in \Gamma.$$

Setting the first derivatives of L_h equal to zero and simplifying, one obtains

$$v_{gh} = (-\lambda \cdot A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*), \quad g \in \Gamma, \quad (12)$$

$$(-\lambda)^{1/2} = \sum_g (N_{gh} \cdot A_{gh})^{1/2} / D_{\Gamma h}, \quad (13)$$

where

$$D_{\Gamma h} = C_h^2 \cdot Y_h^2 \sum_{g \in \Gamma} \left(\frac{1}{v_g^*} \right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^*} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Note that there is no solution to the allocation problem unless $D_{\Gamma h}$ is positive. Substituting (13) into (12) yields

$$v_{gh}^* = (A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*) \cdot \sum_{g \in \Gamma} (N_{gh} \cdot A_{gh})^{1/2} / D_{\Gamma h}. \quad (14)$$

If v_{gh}^* is greater than one for certain gh , the overallocation procedure described above can obviously be employed. Note that (14) also provides the solution for step (ii) of each exact method iteration.

4. EMPIRICAL STUDY

The approximate method serves two purposes. First, it provides a good starting value for the exact method. Second, it may be easier to implement in practice. In this section, we report the results of an empirical comparison using data from the province of Quebec for tax year 1988. Results obtained using the exact method with various starting points, as well as the approximate method, are reported. Since the quantities N_{gh} , Y_h and S_{gh}^2 required by both methods were unknown, estimates based on the data were used.

The size stratification used by the survey, including four take-some strata and one take-all stratum, was employed. Allocations were computed for 64 SIC2 cells (all of the Quebec data excluding a few small SIC2s). The number of sampling fractions determined in these allocations ranged from 8 to 92 with a median of 24. The number of constraints ranged from 9 to 115 with a median of 31. There were 20 SIC2 cells involving more than 35 variables and 18 of these cells also involved more than 50 constraints. A total of 1850 second-phase strata including about 230,000 population units were involved.

The first-phase sampling cost, corresponding to the cost of microfilming or photocopying a tax return at Revenue Canada, sending the information to Statistics Canada and determining an SIC4 code, was set at \$1.40 per unit. The second-phase sampling cost, corresponding to the cost of transcribing values for financial variables, was set at \$7.00. These costs are comparable to those incurred during operation of the actual survey.

Allocations were computed using the exact method with three starting values: I – solution of the approximate method; II – all first-phase sampling fractions set to one with the corresponding conditionally optimal second-phase fractions; and III – a randomly chosen set of feasible first-phase sampling fractions, with the corresponding conditionally optimal second-phase fractions. In addition, the exact method was started at a perturbation of each of these starting values. The perturbed value for the first-phase sampling fraction for size stratum g for starting value I was $v_g^{(0)} = 0.1 + 0.9 \cdot v_g^*$, where v_g^* is the solution of the approximate method. Second-phase sampling fractions were started at values that are optimal, conditional on the perturbed first-phase fractions. Starting value III was perturbed analogously. The perturbed value corresponding to starting value II was $v_{gh}^{(0)} = 0.1 + 0.9 \cdot v_{gh}^{**}$, where v_{gh}^{**} is optimal, conditional on a census at the first phase of sampling. For each starting value, the best result obtained using either the value itself or the corresponding perturbed value was retained. Convergence was declared if the absolute relative change in the cost function between consecutive iterations was less than 10^{-4} . The IMSL implementation of Schittkowski's successive quadratic programming algorithm was used to solve nonlinear programming problems.

Results are reported in Table 2. Total costs for four alternatives are given. In addition, the number of SIC2 cells for which each starting value for the exact method produced better results than alternative starting values is shown. Computing costs are not reported, since they were small enough to be inconsequential.

The results indicate that the approximate solution provided the best starting values for the exact method. Although starting value II produced better results than starting value I for 17 SIC2 cells, the total cost associated with starting value II was higher than the total cost for the approximate method. The exact method performed poorly when starting values were determined by random selection of a feasible set of first-phase sampling fractions.

Table 2
Results for Exact and Approximate Methods

Method	Exact – Starting value			Approximate
	I	II	III	
Total cost (\$)	122779	139347	200998	130228
No. cells with best result*	48	17	1	

* For two cells starting values I and II produced the same result, which had lower cost than the result obtained using starting value III. Consequently, the numbers reported in this row of the table add to 66 rather than 64.

Although the total cost using the exact method with starting value I was only 5.7% lower than the cost of the approximate method, it should be noted that the exact method with starting value I can do no worse than the approximate method. The exact method with starting value I produced better results than the approximate method for 42 cells.

5. CONCLUSION

A sample allocation problem for two-phase survey designs is formulated as a constrained optimization problem in Sections 1 and 2. If the numbers of variables and constraints involved in the problem are small, the solution can be obtained through direct application of numerical methods. However, the direct approach does not work well for large numbers of variables and constraints.

By exploiting the mathematical structure of the problem, it can be divided into two subproblems: the first is a convex programming problem with linear constraints that involves a much smaller number of variables, and the second can be solved without the use of numerical methods. The algorithm proposed in Section 2 consists of iterations between the two subproblems. It is computationally simpler and more effective in practice than the direct approach for problems involving large numbers of variables and constraints. An approximate solution to the sample allocation problem that does not require use of numerical methods is proposed in Section 3. The empirical study in Section 4 shows that it works especially well as a starting value for the algorithm proposed in Section 2.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of Pierre Lavallée, who was the first to derive the expressions for conditionally optimal second-phase sampling fractions involved in the approximate method. Thanks are due to a referee and an associate editor for useful comments. C.F.J. Wu is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- ARMSTRONG, J.B., BLOCK, C., and SRINATH, K.P. (1991). Two-phase sampling of tax records for business surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 228-233.
- BOOTH, G., and SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.
- CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989a). Two-phase sample design for tax data. Unpublished document, Business Survey Methods Division, Statistics Canada.
- CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989b). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.
- IMSL (1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- SCHITTKOWSKI, K. (1985). NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.
- SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.
- SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition? *Journal of the American Statistical Association*, 84, 916-921.
- SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.

The Role of the Interviewer in Survey Participation

MICK P. COUPER and ROBERT M. GROVES¹

ABSTRACT

Using data from a survey of U.S. Census Bureau interviewers, this paper examines whether experienced interviewers achieve higher response rates than inexperienced interviewers, controlling for differences in survey design and attributes of the populations assigned to them. After demonstrating that the relationship is positive and curvilinear, it attempts to explain the mechanisms by which experienced interviewers achieve these rates and elaborate the nature of the relationship. It examines what behaviors and attitudes underlie the higher success, with the hope that they might be instilled in trainees.

KEY WORDS: Interviewers; Nonresponse; Response rates; Survey participation.

1. INTRODUCTION

Survey methodologists have long suspected the interviewer to be an important source of variation in response rates. Indicators of this include observed differences among trainees in the ability to absorb and put into practice the interviewing guidelines, interviewer variation in item missing data rates, individual interviewers' response rates, and the ability of some interviewers to convert the initial refusals of others. However, several of these indicators are affected by the fact that interviewers often do their work in different subpopulations, and thus face different challenges to complete their assignments.

Much of what we believe about the impact of the interviewer on survey participation remains untested or inconclusive. In an oft-cited study, Durbin and Stuart (1951) found experienced interviewers to be "decidedly superior" to student volunteers in terms of response rates. Groves and Fultz (1985) found that novice interviewers (1 to 6 months of tenure) had the highest refusal rates in a telephone survey. In a study cited by Inderfurth (1972), nonresponse rates for Census Bureau interviewers trained in 1962 and 1963 declined steadily over the first months of service, reaching the level of experienced interviewers after 22 months. In contrast, Singer, Frankel and Glassman (1983, p. 74) found the effect of experience on response rates in a telephone survey to be counter-intuitive, that is, more experienced interviewers did **not** achieve higher response rates. They do note, however, that this result is based on only six interviewers. In a study of 16 field interviewers in Sweden, Schyberger (1967) found nonresponse rates to be **higher** for experienced than for newly recruited interviewers. In short, the common belief of experienced interviewers being more successful is not uniformly supported empirically.

This paper examines the role of various interviewer characteristics, particularly experience, in achieving respondent cooperation. It should be noted that the interviewer represents only one part of a large set of factors that can affect survey participation. Such factors include respondent characteristics, the respondent-interviewer interaction, survey design features, and contextual and situational factors. For a review of these factors, see Groves, Cialdini and Couper (1992).

¹ Mick P. Couper and Robert M. Groves, U.S. Bureau of the Census and University of Michigan, Room 2315-3, Bureau of the Census, Washington, DC 20233.

We should also note that different models may be more suitable for different components of nonresponse. For instance, interviewer motivation, tenacity and effort expended may be more important in reducing noncontacts, while persuasion skills play a greater part in the refusal component of nonresponse. The data analyzed here do not permit us to distinguish between these components of nonresponse. This may weaken the explanatory power of the models tested.

In this paper we will address two questions: (a) do experienced interviewers achieve higher response rates? (b) if so, what are the mechanisms underlying the relationship between experience and rates? These questions are important to the survey research community. If the behaviors used by successful experienced interviewers can be taught to inexperienced interviewers, then their success might be transferred to the new recruits. If not, then the value of reducing turnover among experienced interviewers remains high for survey organizations.

2. TOWARD A MODEL OF SURVEY PARTICIPATION

A number of interviewer characteristics can be identified that have a potential impact on survey participation. These are illustrated in Figure 1. The effects of interviewer experience, expectations and behavior on response rates, controlling for assignment area and survey design features, will be explored. Each of the sets of variables will be discussed in turn.

2.1 Interviewer experience

First, interviewers' experience is expected to have a positive effect on the response rates they obtain. This stems from lessons learned through trial and error application of alternative techniques over time, and from alternative training guidelines and experiences on different surveys. Experience thus has two components: length and breadth. Length of experience might be indicated by the number of years a person has worked as an interviewer. One indicator of breadth of experience is the number of different organizations an interviewer has worked for, or the number of different kinds of studies an interviewer has worked on. It is argued that length and breadth of experience both serve to increase the variety of different interviewing situations to which an interviewer is exposed.

We expect the relationship between length of experience (as measured by tenure) and response rates to be curvilinear. Experience in the first few years of interviewing will have a greater impact on response rates than in later years. After a certain point, the number of new situations faced by interviewers declines, and interviewers become comfortable dealing with the wide variety of sample persons and assignment areas they may face. After this, additional years of experience may not produce further gains in response rates.

An alternative hypothesis is that self-selection rather than experience produces higher response rates among interviewers with longer tenure. In other words, it is not that individual interviewers get better over time, but that better interviewers tend to stay, while weaker interviewers leave the job. We believe that a combination of these two factors explains variations in interviewer performance. However, the self-selection hypothesis cannot be tested in a cross-sectional study such as this, and caution must be exercised in drawing inferences from these analyses.

If experienced interviewers achieve higher response rates, we hypothesize that this takes place through the intervening effects of interviewer expectations (*e.g.* confidence) and behavior (*e.g.* effective oral presentation). Note that we posit no direct effect of experience on response rates. In other words, is it possible to identify interviewer attitudes and behaviors that may account for possible differences in response rates?

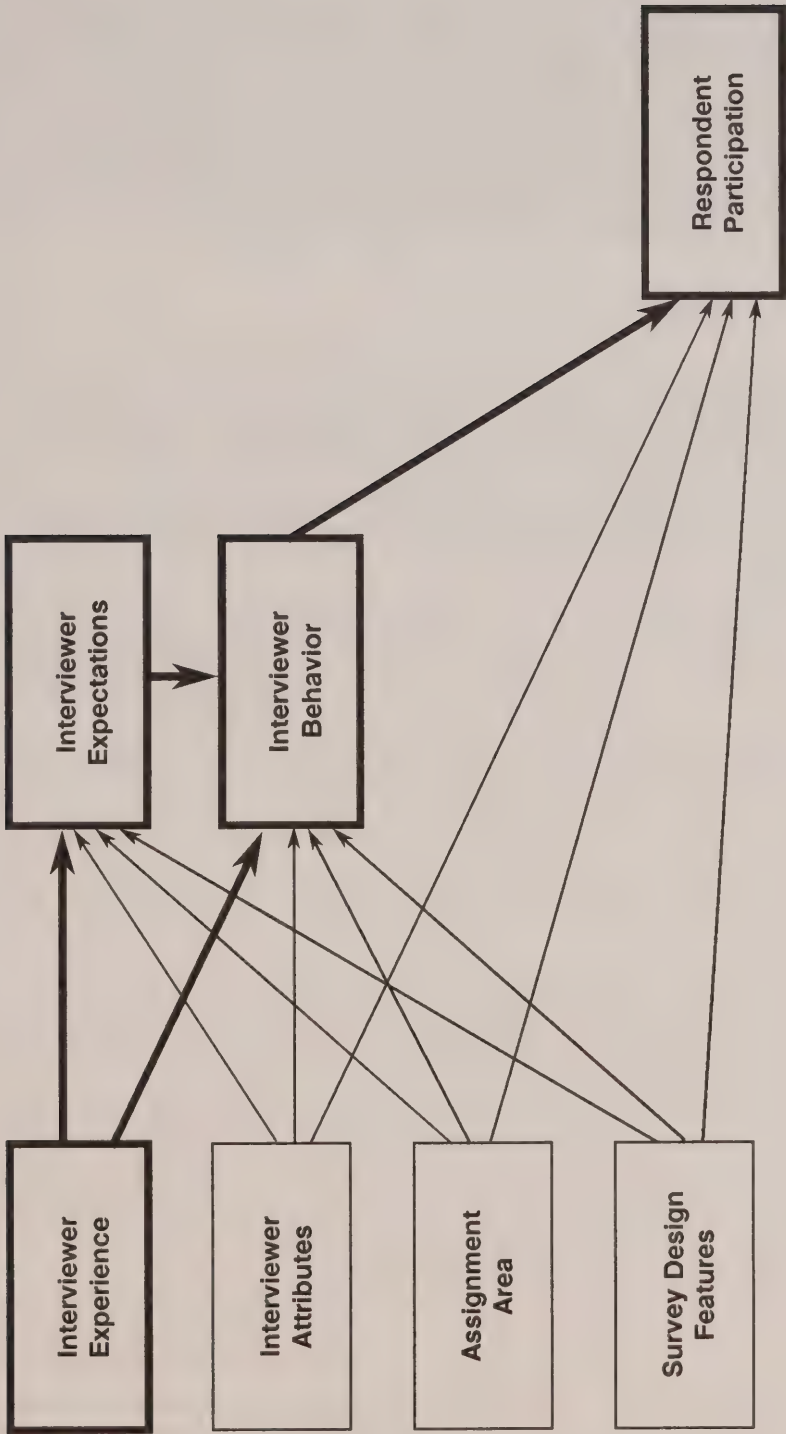


Figure 1. Model of Survey Participation Role of the Interviewer.

2.2 Interviewer expectations

It is hypothesized that positive interviewer expectations lead to higher response rates. Interviewers who have a greater belief in their ability to persuade sample persons to participate, who believe in the legitimacy of the work they are doing, and who are confident that most people agree to participate in surveys, are likely to get higher response rates than those who believe otherwise. This argument has some empirical support in the study by Singer, Frankel and Glassman (1983), in which it was found that interviewers who anticipated prior to the survey that the task of persuading respondents was “moderately easy”, achieved higher response rates than those who believed the task to be “moderately difficult”.

2.3 Interviewer behavior

With regard to interviewer behaviors, we seek to identify the mechanisms by which greater experience and positive expectations translate into higher response rates. The behavior of interviewers in gaining cooperation from sample persons may be likened to that of other “compliance professionals” (such as salespersons, fundraisers, *etc.*). Based on an extensive review of experimental and observational evidence, Cialdini (1984, 1990) identifies six compliance principles used to decide whether to accede to a request. Briefly, these principles are as follows:

- (a) Reciprocation: One should be more willing to comply with a request to the extent that the compliance constitutes the repayment of a perceived gift, favor, or concession.
- (b) Consistency: After committing oneself to a position, one should be more willing to comply with requests for behaviors that are consistent with that position.
- (c) Social validation: One should be more willing to comply with a request to the degree that one believes that similar others would comply with it.
- (d) Authority: One should be more willing to yield to the requests of someone who one perceives as a legitimate authority.
- (e) Scarcity: One should be more willing to comply with requests to secure opportunities that are scarce.
- (f) Liking: One should be more willing to comply with requests of liked others.

We are interested in the extent to which interviewers make use of these principles to persuade sample persons to participate in a survey.

It is argued that interviewers who make appropriate use of each of these strategies are likely to have greater success in persuading reluctant sample persons to participate. However, the use of such techniques indiscriminately in all situations may backfire. For example, the invocation of the authority principle in areas where suspicion of government is high may well have a negative effect on cooperation. The use of these compliance principles may not be universally effective in all situations or for all sample persons.

Thus, it is not just **whether** these techniques are used by interviewers, but also **how** they are used. Two concepts are of interest here. One is the number of different techniques that an interviewer has at his/her disposal, and the second is how appropriately such techniques are applied. The first we will refer to as the “repertoire of techniques” available to the interviewer. A novice interviewer may learn one or two “canned” introductions during training, and use them on all sample persons he/she encounters. In contrast, the experienced interviewer has a wide repertoire of approaches upon which to draw, and can apply them as the situation warrants.

The second concept is that of appropriate application of the skills or techniques at the interviewer's disposal. We refer to this as "tailoring". An interviewer is expected to be an "astute psychological diagnostician" (Cannell 1964), to be able to size up a situation quickly, and apply the appropriate persuasive messages. These skills are gained through experience, either on the job or in life in general. The novice interviewer, with fewer skills and less confidence, may rigidly adhere to a small number of "tried and trusted" approaches. The experienced interviewer is better able to tailor his/her approach to each potential respondent.

It may be that adaptability and appropriate application of persuasive techniques are more critical than the actual behaviors or techniques themselves. If so, it should be possible to develop a more parsimonious model using only the latter concepts and dropping the specific behaviors measured.

2.4 Assignment area

To examine the effect of interviewers on survey participation, we need to take into account the fact that they are assigned different areas to interview. Ideally, the research design would have randomly assigned interviewers to sample areas, removing any statistical confounding between interviewer and population characteristics. Without such randomization, we attempt to specify those population characteristics important to response rate and statistically control for them.

First, the problem of obtaining cooperation from sample persons in inner-city areas is well known (see Steeh 1981, Smith 1983). House and Wolf (1978) found that rising crime rates, particularly in high density urban areas, have been a major deterrent to survey participation, and to trusting and helping behavior in general (Korte and Kerr 1975). We expect this arises both because of residents' reluctance to interact with strangers, and unease among interviewers on entering these neighborhoods.

Turning to characteristics of sample households, household size has been found to correlate positively with response rates (see Gower 1979; Paul and Lawes 1982; Rauta 1985). Single-person households tend to have relatively high refusal rates (see Brown and Bishop 1982; Wilcox 1977). This may be due in part to the large proportion of elderly persons living alone. Families with dependent children, on the other hand, tend to have higher response rates. Lievesley (1988) notes that higher response rates in certain areas of the U.K. may be explained by the high probability of finding someone at home arising from high proportions of children aged 0-4.

The findings on sample person characteristics are somewhat more mixed. A number of researchers (see Brown and Bishop 1982; Hawkins 1975; Herzog and Rogers 1988; Weaver 1975) have found age to be associated with nonresponse. The impact of other sample person characteristics such as race, education, socio-economic status, gender, *etc.* are somewhat inconsistent (see Groves (1989) and Goyder (1987) for reviews of these factors).

2.5 Survey design features

Finally, survey design features (topic, burden, respondent selection rules, *etc.*) are likely to influence a sample person's decision to participate, both directly and in terms of constraints on interviewer expectations and behavior.

2.6 Interaction effects on response rate

We suspect that there may be a number of statistical interaction effects of influences on nonresponse. One question is whether there are some areas (such as high density central city areas) in which interviewer experience is more important than other areas. For example, high density urban areas may be more diverse, requiring greater experience to deal with a greater

variety of different situations. Behavior in areas where the situations presented to interviewers are all very similar could be more easily learned, as fewer persuasion strategies would be needed.

We also suspect that different surveys may obtain varying response rates for different subpopulations as a result of the differential salience of the survey topic to such groups. For example, it may be expected that the National Crime Survey (which focuses on criminal victimization) may get higher response rates in high crime areas than in low crime areas. Similarly, the National Health Interview Survey (which measures health-related activities) may obtain higher response rates in areas with an older than average population. Similar interactions may be expected between the Consumer Expenditure Survey and such variables as average household size and income level.

3. METHOD

3.1 Data collection strategies

The results in this paper are part of a larger study of survey participation in face-to-face surveys in the United States. The first part of the work involved a series of focus groups with interviewers working on a variety of different surveys around the country. The insights gained from these groups led to the development of a structured questionnaire to test some of these hypotheses on a larger audience of interviewers.

The interviewer surveys had the goal of measuring behavioral, experiential and attitudinal influences on levels of cooperation obtained by interviewers. The questionnaire was developed and tested by staff at the Survey Research Center in collaboration with staff from the U.S. Census Bureau.

This questionnaire was administered to U.S. Census Bureau interviewers working on the following three personal visit surveys:

- (a) the Consumer Expenditure Quarterly Survey (CE), sponsored by the Bureau of Labor Statistics;
- (b) the National Health Interview Survey (HIS), sponsored by the National Center for Health Statistics; and
- (c) the National Crime Survey (NCS), sponsored by the Bureau of Justice Statistics.

The questionnaire was mailed in February, 1990, to Census Bureau interviewers working on these three surveys. All interviewers were paid their normal salary rate for completing the questionnaire (most were paid for an hour of their time). In an effort to seek candid responses and eliminate the threat of supervisory intervention, interviewers were assured that their individual responses would not be seen by or discussed with any of their supervisors, and that the results would be reported only as statistical totals.

Questionnaires were mailed back to the central office. Reminder letters and telephone calls were used to increase the response rate. A total of 1,013 completed questionnaires were received, representing a response rate of 97.1%. A number of questionnaires were excluded from the analyses reported here. All supervisory interviewers (256) were excluded. These people often have no regular assignments of their own, and typically work on a number of different surveys. They are often used for refusal conversion, or to "clean up" otherwise incomplete assignments. With supervisory interviewers excluded, transfer of assignments from one interviewer to another on these surveys is rare. For purposes of calculating interviewer-level response rates, each nonresponse case was counted against the original interviewer, regardless of whether it was later converted by another. In addition, those interviewers who started work during the period

in which the interviewer survey was administered, and for whom no historical response rate information was available, were also excluded (46 interviewers). This left a total of 711 interviewers, 207 from CE, 139 from HIS and 365 from NCS. The numbers of cases included in the analyses may be further reduced due to missing data on certain variables.

3.2 Data structure

In addition to the questionnaire responses, other variables were added to the data file. These included a set of variables to represent each interviewer's assignment area. Typically, the primary sampling unit (PSU) in which an interviewer works consists of one or more coterminous counties. County-level data were extracted from the County and City Data Book (Bureau of the Census 1988), aggregated to the PSU level, and attached to the interviewer records. Note that these variables can only reflect gross differences in assignment area and cannot, for example, distinguish between central city and suburban areas.

The date each interviewer was hired by the Census Bureau was obtained from administrative records to create a variable to serve as a measure of tenure. Although it does not indicate length of experience on a particular survey, it does reflect the length of time an interviewer was employed by the Census Bureau.

A major drawback of this study is that it was not possible to obtain measures of race, age, gender, or other demographic attributes of the interviewer. Confidentiality restrictions prevented access of personnel records for this information, nor could these be asked in the interviewer questionnaire.

3.3 Analytic plan

Three different surveys are represented in the data set. Instead of introducing control variables measuring key design features of the surveys, dummy variable indicators of the survey were used to control on important design differences among them.

The dependent variable is aggregate response rate for the six month period, October 1989, through March 1990. It was not possible to obtain interviewer-level data on the components of nonresponse (particularly refusals) for this period. These rates thus do not distinguish between noncontact and refusal components of nonresponse. Hence, it should be noted that the analyses reported here are based on interviewer-level **response** rates rather than **refusal** rates.

The nonresponse rates for the three surveys for 1990 (based on national sample totals) are presented in Table 1.

Refusals as a proportion of total nonresponse varies from 87% for CE to 52% for NCS. We suspect that different sets of factors operate to affect these two components of nonresponse. Ideally, separate models would be fitted for each component, but this was not possible given the current data. To the extent that factors affecting refusals are different from those affecting other components of nonresponse (such as noncontacts), the results will be confounded (see Lievesley 1988). It can also be seen that nonresponse rates for these three surveys are low to begin with. This may further restrict the ability of these models to explain differences among interviewers.

Given that the size of the interviewer assignments vary (and hence affect the variance of the measured individual response rates), we used weighted least squares (WLS) with assignment size as the weight. Comparisons of the WLS results with those using ordinary least squares (OLS) solutions were made, and it was found that WLS reduces the size of the coefficients marginally, but does not affect the sign or relative strength of the coefficients. All the analyses reported here are based on the WLS solutions.

Table 1
1990 Nonresponse Rates for Three Surveys

Survey	Nonresponse rate	Refusal rate
	%	%
Consumer Expenditure Survey	13.4	11.6
Health Interview Survey	4.5	2.8
National Crime Survey	3.1	1.6

A series of tests were performed to determine the appropriateness of the models specified. A number of outliers in the dependent variable were detected. However, removal of these outliers had little or no effect on the results obtained, and they were therefore retained in all analyses. Tests of the normality assumption were also conducted. The normal probability plots show that the residuals from these models do not differ markedly from a normal distribution.

It is hypothesized that the effect of tenure on response rate is greater in the first few years. The tenure variable is transformed (the natural log is used) to reflect this. The transformed variable indeed produced an improvement in fit over the linear tenure variable.

A more detailed description of the variables used in these analyses can be found in Appendix A.

4. LIMITATIONS

Before describing the analyses, it is important to note some of the limitations of these data. First, these findings refer only to interviewers working on three ongoing national surveys at the Census Bureau at the time at which the interviewer survey was conducted. It is not possible to generalize to other face-to-face or telephone surveys conducted by academic or private sector organizations.

Furthermore, the data are cross-sectional in nature. Cohort and period effects are confounded with the effects of experience. That is, any observed response rate differences by interviewer experience may be due to changes in the quality of interviewers hired over time, in the effectiveness of interviewer training over time, or in differential turnover by interviewer quality. Hypotheses can be constructed to support both positive and negative effects of these factors on response rates. Hence, the measured impact of interviewer experience on response rates is a complex combination of these factors. Longitudinal measurement of interviewers is needed to disentangle these effects.

Interviewers are not randomly assigned to areas. Although we have attempted to control for a number of characteristics of assignment area that may impact on response rates, there may be many other factors that could explain differences in response rates across assignment area. Further, we are limited to weak controls, on attributes of counties and groups of counties, not on attributes of specific assignment areas within counties given to interviewers. A hierarchical analysis containing data on individual respondents and interviewers assigned to them would improve these control factors.

Finally, the dependent variable was measured for a time period up to and including the administration of the interviewer questionnaire. More recent response rate data were not available at the time. Given that behaviors and expectations were not measured before the response rates were obtained, caution should be exercised in attributing causality.

Despite these limitations, these data provide us with the opportunity to test prevailing beliefs about the role of interviewer experience in response rates, and to explore the role of interviewer expectations and behavior in face-to-face surveys.

5. RESULTS

First, we measured the impact of experience, controlling for characteristics of assignment areas and dummy variables for the surveys (Model 1 in Table 2). Let us first examine the coefficients of the control variables. With few exceptions, most of the assignment area variables have a significant impact on response rates. Both population density and crime rate act as expected, with lower response rates being obtained in high crime, high density areas. The negative effect of household size is contrary to expectation. This may be explained in part by the fact that these surveys all collect information from or about **all** adult household members, thereby increasing the reporting burden for large households. This is contrary to many surveys where a single adult is selected from each household. The effect of age is as hypothesized, with response rates tending to be lower (but not significantly so) in areas with larger proportions of persons over 65, but higher in areas with many households who have young children.

The large effects for the two survey variables (relative to the omitted category of the Consumer Expenditure Survey) reflect differences in the mean response rates for these three surveys. Such differences can be attributed to a host of survey design differences (length of the interview, respondent selection rules, panel versus cross-sectional designs, content of the questionnaires, *etc.*) that are beyond the scope of this paper. Nevertheless, it is clearly necessary to control for these differences.

Now, let us examine the measured effect of experience, given these control variables. It can be seen that tenure has a strong positive effect on response rates, even when controlling for the nature of the area to which an interviewer is assigned. This appears to confirm prevailing beliefs about the role of interviewer experience. Interviewer differences in response rates appear to be more than simply artifacts of differences in the areas to which they are assigned, and experience plays a key role in such interviewer differences.

The inclusion of an indicator for breadth of experience was also tested, but found to have no significant effect in the presence of the remaining variables. It thus appears that, for Census Bureau interviewers at least, experience working for other survey organizations does not appear to have any marginal impact on response rates over and above that of tenure.

Does tenure have a differential impact on response rates in different assignment areas? Model 2 in Table 2 includes an interaction term between the log of tenure and population density. An additional interaction term between tenure and crime rate was also tested, but this coefficient was found to be insignificant, and the interaction had little impact on remaining elements of the model. The interaction term in Model 2 is statistically significant, but the sign is opposite to that expected. We hypothesized that experience would have a greater impact in high density areas, but this does not appear to be the case. An alternative explanation may be a "burnout effect". More experienced interviewers in high density urban areas may be losing their enthusiasm sooner than experienced interviewers in less stressful rural areas, and this contributes to lower response rates. Interviewer burnout may be one factor contributing to higher turnover rates in the large metropolitan areas.

Table 2
Results of WLS Regression Analyses of NCS, HIS, CE Interviewer-Level Response Rates

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Intercept	96.94	(3.19)	96.21	(5.39)	94.95	(3.25)	93.44	(3.35)
Assignment area:								
Population density	−0.00017**	(0.000023)	−0.000078*	(0.000038)	−0.000084*	(0.000038)	−0.000071	(0.000038)
Crime rate	−0.00024**	(0.000055)	−0.00021**	(0.000055)	−0.00023**	(0.000056)	−0.00022**	(0.000056)
Percent 65 or older	−0.057	(0.051)	−0.054	(0.050)	−0.061	(0.051)	−0.061	(0.052)
Percent under 5	0.41*	(0.16)	0.37*	(0.16)	0.29	(0.17)	0.35*	(0.17)
Household size	−3.20*	(1.70)	−2.92*	(1.24)	−2.88*	(1.26)	−3.09*	(1.27)
Survey Indicators:								
NCS ¹	6.72**	(0.40)	6.67**	(0.40)	6.68**	(0.41)	6.55**	(0.42)
HIS ¹	5.65**	(0.46)	5.63**	(0.46)	5.64**	(0.47)	5.65**	(0.48)
Interviewer experience:								
Log (tenure)	0.62**	(0.14)	0.74**	(0.14)	0.69**	(0.15)	0.72**	(0.15)
Log (tenure) × density			−0.00010**	(0.000032)	−0.00011**	(0.000032)	−0.00011**	(0.000032)
Interviewer expectations:								
Confidentiality								
Rate/quality					0.61	(0.37)	0.59	(0.37)
Efficacy					0.046	(0.40)	−0.00073	(0.41)
					0.55**	(0.15)	0.53**	(0.15)
Interview behaviors:								
Authority							0.14**	(0.055)
Reciprocation							0.67*	(0.29)
Social proof							0.18	(0.32)
Saliency							−0.19	(0.33)
Scarcity							−0.66*	(0.29)
Consistency							−0.21	(0.29)
Repertoire							−0.0068	(0.065)
Tailoring							−0.042	(0.054)
Adjusted R ²	0.3553	(679)	0.3640	(679)	0.3784	(645)	0.3873	(639)
(n)								

** $p < .01$
* $p < .05$
1 CE Interviewers are the omitted category.

Interactions between the three surveys and various assignment characteristics were also tested. None of these appear to have any noticeable effect in these models, and are not discussed further. As a further test for the presence of additional interactions involving the survey variables, separate models were fitted for each of the three surveys. The models obtained are essentially the same for each of the three surveys examined. Thus, although the level of response differs across the three surveys, the **relative** impact of tenure on response rates appears to be the same.

Given that it appears that experienced interviewers achieve higher response rates regardless of the areas to which they are assigned, we can proceed to address the question of **how** experience impacts on levels of cooperation. What makes a more experienced interviewer better at gaining cooperation from respondents?

The first step involves the addition of interviewer expectation variables to Model 2. The results are presented as Model 3 in Table 2. All three expectation variables act in the expected direction, although only one achieves statistical significance at traditional levels. It appears that those interviewers who have a greater belief in their ability to convince reluctant respondents to participate, actually achieve higher response rates.

It should be cautioned that the causal link between expectations and response rates cannot be established in a cross-sectional study such as this. It may be that greater success leads to greater expectations of future success, rather than the other way around. This interpretation opposes the hope that instilling a greater sense of self-efficacy in interviewers will produce higher levels of response. Nevertheless, this finding is an intriguing one that demands further attention.

The next step was to add the set of interviewer behaviors into the model. The results can be seen in Model 4 in Table 2. Two things can be noted about these results. First, the inclusion of this set of interviewer behaviors failed to explain away the effect of tenure. In fact, the coefficient for tenure is hardly affected by the addition of either the expectation variables or the behavior variables.

Second, the results for the specific behaviors are somewhat mixed. It was expected that the coefficients for all the behavior variables would be positive. This is not the case. The results for authority and reciprocity indicate that interviewers who use these techniques achieve higher response rates. In contrast, use of the scarcity principle appears to have the opposite effect. Pressure on a respondent to meet certain deadlines may well backfire. The remainder of the behavior variables do not appear to have a significant effect on the response rates attained by Census Bureau interviewers.

It was suggested earlier that a reduced model, using only repertoire and tailoring, should be considered. In Table 2 it was seen that these two variables do not have significant effects in the presence of the other behavior variables. Even after removing the other behavior variables from the model, repertoire and tailoring still have little impact on response rates. Thus, the argument that the way interviewers use various compliance techniques are more important than the actual behaviors themselves gains little empirical support from these data. However, the measures of these two concepts may be weak, and a better test of their role should be done at the contact-level of analysis.

6. DISCUSSION

This paper set out to measure whether experienced interviewers achieve higher response rates than inexperienced interviewers. It found they do. It then tried to explain why they do. It largely failed. One reason may be that the model is incorrect. However, continued discussions with interviewers and supervisory staff lead us to believe that this theoretical formulation has some merit.

Four explanations can be posited. First, the model is being tested at the wrong level of aggregation. Although the questionnaire focused on what interviewers usually or typically do, we are more interested in how they act in specific situations. A more appropriate test of these ideas should be conducted at the contact or household level. Second, the measurement of various concepts may be inadequate. Improvements in the translation of concepts from the compliance literature into specific interviewer behaviors may be made. Third, it should again be noted that these models deal with response rates not refusal rates. It may be that certain behaviors are more appropriately directed at persuading sample persons to participate (aimed at reducing refusals), while others may serve more to gain access to sample persons (the non-contact portion of nonresponse). Separate models for these two processes could not be developed here. Finally, other unmeasured characteristics of interviewers (appearance, voice quality, dress, *etc.*) may also play a role in influencing the respondent's decision.

These possible shortcomings do not negate the role of these behaviors in affecting response rates. Rather, the findings suggest further research and analysis to explore the relationships between specific behaviors and their application on the one hand, and interviewer-level response rates on the other. We feel that this line of inquiry has merit, and are working toward a fuller understanding of the role of interviewer experience, expectations and behavior in survey participation.

ACKNOWLEDGEMENTS

This work was supported by the Bureau of the Census, Bureau of Labor Statistics, Bureau of Justice Statistics, and the National Center for Health Statistics. Views expressed are those of the authors and do not necessarily reflect those of the Bureau of the Census or any other organization. The authors wish to thank Lorraine McCall for her assistance with this research. The reviewers are also thanked for their valuable suggestions.

APPENDIX A

VARIABLES USED IN ANALYSES

The creation of the variables used in the analyses are summarized here. Copies of the questionnaire can be obtained from the authors.

Dependent variable

Response rate: This is the response rate obtained by each interviewer for the six-month period in question, expressed as a percentage.

Assignment area

Population density: Population density (persons per square mile).

Crime rate: Crime rate (crimes per 100,000 population).

Percent 65 or older: Percentage of population 65 years of age and older.

Percent under 5: Percentage of population under 5 years of age.

Household size: Average household size.

Survey

Set of dummies to indicate which survey each interviewer works on:

HIS: Does interviewer work on the Health Interview Survey.

1 = Yes

0 = No

NCS: Does interviewer work on the National Crime Survey.

1 = Yes

0 = No

CE: (the Consumer Expenditure Survey) is thus the omitted category.

Interviewer experience

Tenure: Measured in days of service employed at the Census Bureau as an interviewer, rescaled to fractional years.

Breadth of experience: A count of the number of different survey organizations for which an interviewer has worked.

Interviewer expectations

Confidentiality: Interviewers were asked whether they thought there were any situation under which the Census Bureau would give individual survey response to any of a number of agencies (FBI, CIA, INS, IRS, state and local government agencies).

1 = High confidentiality belief (Census Bureau would not give responses to any of these agencies).

0 = Low confidentiality belief (Census Bureau would give responses to one or more of the agencies).

Rate/quality: Trade-off between response rate and data quality. Which one of the following statements comes closest to how you feel as an interviewer:

1 = It's better to persuade a reluctant respondent to participate than to accept a refusal.

0 = It's better to accept a refusal from a reluctant respondent.

Efficacy: Interviewers were asked the extent to which they agreed or disagreed with the following statement: With enough effort, I can convince even the most reluctant respondent to participate.

Four-point ordinal scale, 1 = strongly disagree, 4 = strongly agree. High score indicates greater belief in self-efficacy.

Interviewer behaviors

Authority: Interviewers were asked how often they left various materials (request for appointment, copy of the advance letter, *etc.*) at respondents' home when they found no-one at home. The responses to these questions were combined to form a scale of frequency of use of these authority-enhancing materials. High score indicates greater use of authority.

- Reciprocation:** How often do you make a point of complimenting something about respondent's home or personal appearance?
 1 = Always, sometimes
 0 = Rarely, never
- Social proof:** How often do you say "Most people enjoy doing the interview"?
 1 = Always, sometimes
 0 = Rarely, never
- Saliency:** How often do you explain to respondents how the survey results could affect them personally?
 1 = Always, sometimes
 0 = Rarely, never
- Scarcity:** How often do you tell a respondent that the interview must be completed by a certain date?
 1 = Always, sometimes
 0 = Rarely, never
- Consistency:** Before a respondent has shown any sign of cooperating, how often do you begin asking the survey questions?
 1 = Always, sometimes
 0 = Rarely, never
- Repertoire:** In an open-ended question, interviewers were asked to list all things they usually do to persuade reluctant respondent to participate. A count of the number of distinct things mentioned serves as an indicator of the repertoire of techniques available.
- Tailoring:** In a series of 15 behavior items, interviewers responded whether they always, sometimes, rarely or never performed such behavior. An indicator of tailoring in the application of various persuasion techniques is obtained by counting the number of times an interviewer used the middle categories (sometimes or rarely) to these questions. A high score indicates greater use of tailoring.

REFERENCES

- BROWN, P.R., and BISHOP, G.F. (1982). Who refuses and resists in telephone surveys? Some new evidence. Paper presented at the MAPOR Annual Conference.
- CANNELL, C.F. (1964). Factors affecting the refusal rate in interviewing. Ann Arbor: Survey Research Center (unpublished working paper).
- CIALDINI, R.B. (1984). *Influence; The New Psychology of Modern Persuasion*. New York: Quill.
- CIALDINI, R.B. (1990). Deriving psychological concepts relevant to survey participation from the literatures on compliance, helping, and persuasion. Paper presented at the Workshop on Household Survey Nonresponse, Stockholm.
- DURBIN, J., and STUART, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *Journal of the Royal Statistical Society, Series A*, 114, 163-206.

- GOWER, A.R. (1979). Non-response in the Canadian Labour Force Survey. *Survey Methodology*, 5, 29-58.
- GOYDER, J. (1987). *The Silent Minority; Nonrespondents on Sample Surveys*. Boulder, CO: Westview Press.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.
- GROVES, R.M., CIALDINI, R.B., and COUPER, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly* (forthcoming).
- GROVES, R.M., and FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.
- GROVES, R.M., and KAHN, R.L. (1979). *Surveys by Telephone*. New York: Academic Press.
- HAWKINS, D.F. (1975). Estimation of nonresponse bias. *Sociological Methods and Research*, 3, 461-488.
- HERZOG, A.R., and RODGERS, W.L. (1988). Age and response rates to interview sample surveys. *Journals of Gerontology*, 43, S200-S205.
- HOUSE, J.S., and WOLF, S. (1978). Effects of urban residence on interpersonal trust and helping behavior. *Journal of Personality and Social Psychology*, 36, 1029-1043.
- INDERFURTH, G.P. (1972). Investigation of Census Bureau interviewer characteristics, performance and attitudes: A summary. U.S. Bureau of the Census: Working Paper 34.
- KORTE, C., and KERR, N. (1975). Responses to altruistic opportunities in urban and nonurban settings. *Journal of Social Psychology*, 95, 183-184.
- LIEVESLEY, D. (1988). Unit non-response in interview surveys. London: Social and Community Planning Research (unpublished working paper).
- PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.
- RAUTA, I. (1985). A comparison of the census characteristics of respondents and nonrespondents to the 1981 General Household Survey (GHS). *Statistical News*, 71, 12-15.
- SCHYBERGER, B.W. (1967). A study of interviewer behavior. *Journal of Marketing Research*, 4, 32-35.
- SINGER, E., FRANKEL, M.R., and GLASSMAN, M.B. (1983). The effect of interviewer characteristics and expectation on response. *Public Opinion Quarterly*, 47, 68-83.
- SMITH, T.W. (1983). The hidden 25 percent: An analysis of nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly*, 47, 386-404.
- STEEH, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.
- U.S. BUREAU OF THE CENSUS (1988). *County and City Data Book, 1988*. U.S. Government Printing Office.
- WEAVER, C.N., HOLMES, S.L., and GLENN, N.D. (1975). Some characteristics of inaccessible respondents in a telephone survey. *Journal of Applied Psychology*, 60, 260-262.
- WILCOX, J.B. (1977). The interaction of refusal and not-at-home sources of nonresponse bias. *Journal of Marketing Research*, 14, 592-597.

A Multivariate Procedure Towards Composite Estimation of Consumer Expenditure for the U.S. Consumer Price Index Numbers

P. LAHIRI and WENYU WANG¹

ABSTRACT

We consider the problem of estimating the “cost weights” and “relative importances” of different item strata for the local market basket areas. The estimation of these parameters is needed to construct the U.S. Consumer Price Index Numbers. We use multivariate models to construct composite estimators which combine information from relevant sources. The mean squared errors (MSE) of the proposed and the existing estimators are estimated using the repeated half samples available from the survey. Based on our numerical results, the proposed estimators seem to be superior to the existing estimators.

KEY WORDS: Consumer expenditure; Composite estimation; Consumer Price Index; Cost weight; Diary survey; Half sample; Laspeyres Index; Mean squared error; Synthetic estimation.

1. INTRODUCTION

The U.S. Consumer Price Index (CPI) is an indicator of price changes for a set of items, goods and services, whose quantity and quality are fixed over a period of time. The U.S. Bureau of Labor Statistics (BLS) computes a number of consumer price indices each month for various geographical areas, consumer units and item classification (*vide* BLS Handbook of Methods 1988).

The smallest group of item classification for which the BLS computes the CPI is known as an “item stratum”. It is a prespecified set of consumer goods and services, *e.g.*, fresh whole milk, which can be purchased in the retail market during a “base period” by a specified set of consumer units. A consumer unit may consist of all members of a particular household related by blood, marriage, adoption, or other legal arrangements. A number of item strata constitutes an expenditure class (*e.g.*, dairy products).

The U.S. is divided into eight major areas for sampling purposes. A major area may be either “self-representing” or “non-self-representing” and belongs to one of the four regions (Northeast, Midwest, South and West). A self-representing area consists of all large cities within a region. A non-self-representing area generally consists of a county or a group of contiguous counties. For publication purposes, a major area is further divided into a number of “market basket areas” or “publication areas”.

The Laspeyres formula used by the BLS to compute the CPI for a given area and an expenditure class (say, E) is defined below. Let

P_{it} = the average price of all items in the i th item stratum at time t ($t = 0, T$),

Q_{i0} = the quantity of all items in the i th item stratum purchased at time $t = 0$ (base period).

¹ P. Lahiri, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588 0323, USA. Wenyu Wang, SUNY Health Science Center at Brooklyn, Box 1203, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.

Then the Laspeyres index at time $t = T$ is given by

$$\begin{aligned}
 I_T &= \frac{\sum_{i \in E} Q_{i0} P_{iT}}{\sum_{i \in E} Q_{i0} P_{i0}} \\
 &= \frac{\sum_{i \in E} C_i (P_{iT}/P_{i0})}{\sum_{i \in E} C_i} \\
 &= \sum_{i \in E} R_i (P_{iT}/P_{i0}),
 \end{aligned}$$

where

$C_i = Q_{i0} P_{i0}$ = total expenditure for all items in the i th item stratum at $t = 0$,

$R_i = C_i / \sum_{i \in E} C_i$ = proportion of total expenditure spent on the i th item stratum at $t = 0$.

The quantities C_i and R_i are referred to as the “cost weight” and “relative importance” of the i th item stratum within the expenditure class, E .

The Bureau of Labor Statistics computes the consumer price indices using data from the U.S. Consumer Expenditure Survey (CES). The survey has two different components – Diary survey and Interview survey, each having separate sampling schemes and questionnaires. In this paper we consider data from the Diary survey only. The sampling design selects all the primary stage units (PSU’s) within a particular self-representing area with certainty. But only a sample of PSU’s is selected for a particular non-self-representing area according to a probability sampling scheme. From each selected PSU, a sample of consumer units (CU’s) is selected again using some probability sampling design. Each respondent keeps a diary of expenditures on various items for two consecutive 1-week periods. For a detailed account on the CPI and CES, the reader is referred to the BLS Handbook of Methods (1988).

The efficiency of the traditional sample survey estimators of the cost weight and relative importance of an item stratum at the publication area level is generally very low compared to their efficiency at a larger area (*e.g.*, major area) level. This is due to the fact that only a few consumer units are available from a given publication area. Thus, there is a need to improve the traditional estimator by borrowing strength from related resources. Marks (1978) and Cohen and Sommers (1984) considered certain composite estimators which pool information from related areas. Ghosh and Sohn (1990) obtained composite estimators of the cost weight and relative importance using an empirical Bayes approach.

The current procedure used by the Bureau of Labor Statistics consists of several steps. First composite estimators of the relative importances are obtained using a method suggested by Cohen and Sommers (1984). The estimators of the cost weights are then obtained from these estimators of the relative importances using an iterated “raking” procedure. The final estimates of the cost weights for the entire expenditure class and for the major area are identical to the corresponding preliminary estimates. One reason for ensuring this “data consistency” by raking may be due to the fact that the performances of the preliminary estimators are generally satisfactory at a higher level of aggregation compared to their performances at a lower level. At the last step, the final estimators of the relative importances are obtained directly from the final cost weight estimators by division.

Unlike earlier authors, we use the correlations between the item strata in proposing our composite estimators in Section 2. The shrinkage factor of the composite estimator obtained by minimizing the mean squared error within an appropriate class of estimators involves some unknown parameters. These unknown parameters are estimated using the balanced repeated replications available from the survey. The estimator proposed by Cohen and Sommers (1984) turns out to be a special case of our estimator if one assumes that the preliminary estimators are all uncorrelated.

In Section 2 we concentrate our attention to the estimation of the cost weight of an item stratum for a publication area. However, we can obtain estimators of the cost weights at a higher level of aggregation (*e.g.*, expenditure class for a publication area, *etc.*) by appropriate summation. From our study, it turns out that in terms of the mean squared error criterion these estimators always perform better than the corresponding preliminary estimators and hence better than the BLS estimators (note that due to the raking procedure the BLS estimators are identical to the preliminary estimators at higher levels of aggregation).

In Section 3 we propose a composite estimator of relative importance of an item stratum at the publication area level. Instead of using the preliminary estimators of the cost weights we use the preliminary estimators of the relative importances for all the item strata belonging to the expenditure class under consideration. The preliminary estimators of relative importances of all the item strata within an expenditure class add up to unity. Thus, the variance covariance matrix of the preliminary estimators is singular and this makes the problem different from the problem of estimation of the cost weights. Our procedure deletes one item stratum in an optimal manner and thus avoids the problem of singularity of the variance covariance matrix of the preliminary estimators. Our numerical results show that in terms of the mean squared error criterion the proposed estimator is always the best among all the rival estimators considered.

In Section 4, we present all the numerical results. We have evaluated different estimators of the cost weight and relative importance based on estimated mean squared error obtained by using the balanced repeated half samples (see McCarthy 1969, Ghosh and Sohn 1990). Based on our results, the proposed estimators seem to be superior to all the rival estimators considered in the paper.

2. ESTIMATION OF THE COST WEIGHT

Let X_{ijl} be the average of two consecutive weeks of expenditure for all the items in the i th item stratum by the l th consumer unit belonging to the j th publication area within a particular major area ($i = 1, \dots, I; j = 1, \dots, m; l = 1, \dots, n_j$). Let W_{jl} be the sampling weight attached to the l th consumer unit in the j th publication area ($j = 1, \dots, m; l = 1, \dots, n_j$). This represents a number of consumer units in the population and is obtained by the Census Bureau using a complex procedure which takes into account various factors such as inclusion probabilities, nonresponse, *etc.* In this section, we consider estimation of θ_{ij} , the true average weekly expenditure per consumer unit for the i th item stratum and j th publication area. The cost weight is simply defined as $N_j\theta_{ij}$, where N_j denotes the total number of consumer units in the j th publication area. The preliminary estimator of θ_{ij} is given by

$$Y_{ij} = \sum_{l=1}^{n_j} W_{jl} X_{ijl} / \sum_{l=1}^{n_j} W_{jl}, \quad (i = 1, \dots, I; j = 1, \dots, m). \quad (2.1)$$

Similarly, the corresponding estimator for the major area is given by

$$Y_{i.} = \sum_{j=1}^m \sum_{l=1}^{n_j} w_{jl} X_{ijl} / \sum_{j=1}^m \sum_{l=1}^{n_j} w_{jl}. \quad (2.2)$$

The variability of $Y_{i.}$ is much lower than that of Y_{ij} . Thus, a composite estimator of θ_{ij} which increases the precision is needed. Let $Y_j = (Y_{1j}, \dots, Y_{Ij})'$ and $\theta_j = (\theta_{1j}, \dots, \theta_{Ij})'$, $j = 1, \dots, m$. Let V_j be the true variance covariance matrix of Y_j , ($j = 1, \dots, m$). Under a synthetic assumption, i.e., $\theta_j = \mu$, a $I \times 1$ column vector, ($j = 1, \dots, m$), the best estimator of θ_j is given by

$$\tilde{\mu} = \left(\sum_{j=1}^m V_j^{-1} \right)^{-1} \sum_{j=1}^m V_j^{-1} Y_j, \quad (2.3)$$

which is obtained by minimizing $\sum_{j=1}^m (Y_j - \mu)' V_j^{-1} (Y_j - \mu)$ with respect to μ . The synthetic assumption, however, is hardly satisfied. In the other extreme when there is absolutely no similarity between the θ_j 's, it is appropriate to take Y_j as an estimator of θ_j . When the real situation is in between these two extremes one may take a composite estimator given by

$$\hat{\theta}_{ij}(a_{ij}) = (1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu}, \quad (2.4)$$

where a_{ij} 's are constants ($0 \leq a_{ij} \leq 1$), e_i is a $I \times 1$ column vector having 1 for the i th elements and 0 for the others.

We obtain a_{ij} by minimizing the mean squared error

$$E[\{(1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu} - \theta_{ij}\}^2 \mid \theta_{ij}] \quad (2.5)$$

with respect to a_{ij} . The optimal choice is given by

$$\tilde{a}_{ij} = \frac{e_i' \left[V_j - \left(\sum_{j=1}^m V_j^{-1} \right)^{-1} \right] e_i}{E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \dots, m]}. \quad (2.6)$$

Thus, the optimal estimator of θ_{ij} in the class described by (2.4) is given by

$$\tilde{\theta}_{ij} = (1 - \tilde{a}_{ij}) Y_{ij} + \tilde{a}_{ij} e_i' \tilde{\mu}. \quad (2.7)$$

Remark 1: In the derivation of the optimal estimator $\tilde{\theta}_{ij}$, the quantities V_j , ($j = 1, \dots, m$) and $E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \dots, m]$ are assumed to be fixed and known.

Remark 2: The estimator proposed by Cohen and Sommers (1984) can be obtained from $\tilde{\theta}_{ij}$ as a special case when

$$V_j = \left(\sum_{l=1}^{n_j} w_{jl} \right)^{-1} \text{Diag}(\sigma_1^2, \dots, \sigma_I^2).$$

Note that according to their assumption the correlation between any two item strata is zero which appears to be very restrictive from our study.

Remark 3: Note that using a familiar matrix inversion result (see Rao 1973),

$$V_j - \left(\sum_{j=1}^m V_j^{-1} \right)^{-1} = V_j \left[V_j + \left(\sum_{s \neq j} V_s^{-1} \right)^{-1} \right]^{-1} V_j$$

which is positive definite. Also,

$$E[(Y_{ij} - e_i' \tilde{\mu})^2 | \theta_j, j = 1, \dots, m] = e_i' V_j \left[V_j + \left(\sum_{s \neq j} V_s^{-1} \right)^{-1} \right]^{-1} V_j e_i \\ + \left[\theta_{ij} - e_i' \left(\sum_{j=1}^m V_j^{-1} \right)^{-1} \left(\sum_{j=1}^m V_j^{-1} \theta_j \right) \right]^2.$$

Also, when $\theta_j = \mu$, one gets $\tilde{a}_{ij} = 1$ and thus $\tilde{\theta}_{ij} = e_i' \tilde{\mu}$. Otherwise the size of the shrinkage factor depends on the size of

$$\left[\theta_{ij} - e_i' \left(\sum_{j=1}^m V_j^{-1} \right)^{-1} \left(\sum_{j=1}^m V_j^{-1} \theta_j \right) \right]^2.$$

The larger the distance of θ_{ij} from $e_i' (\sum_{j=1}^m V_j^{-1})^{-1} (\sum_{j=1}^m V_j^{-1} \theta_j)$ the smaller is the size of \tilde{a}_{ij} . This means that if a particular area is very different from the general nature of all the areas then our procedure will give less weight on the synthetic part of the estimator. This explains the great deal of variation of the shrinkage factors in Table 1.

We shall estimate \tilde{a}_{ij} using the 20 balanced repeated half samples available from the survey. Let $w_{jl}^{(k)}$ denote the weight assigned to the l th consumer unit of the j th area for the k th replication ($j = 1, \dots, m$; $l = 1, \dots, n_j$; $k = 1, \dots, 20$). These replicated weights are constructed by the Census Bureau using a complex procedure. For any replication, approximately half the consumer units receive zero weights and the remaining consumer units receive positive weights.

Table 1
Shrinkage Factors \hat{a}_{ij} in West Non-Self-Representing Area

$i \quad j$	1	2	2
1	0.8479225	0.7057626	0.9214804
2	0.8434894	0.5692695	0.8092725
3	0.0969009	0.0786758	0.6953904
4	0.4446537	0.5444809	1
5	0.6999551	0.3460123	0.5487382
6	0.0318442	0.4981756	0.2598752

Define

$$\hat{a}_{ij}^* = \frac{e_i' \left[\hat{V}_j - \left[\sum_{j=1}^m \hat{V}_j^{-1} \right]^{-1} \right] e_i}{\frac{1}{20} \sum_{k=1}^{20} [Y_{ij}^{(k)} - e_i' \hat{\mu}^{(k)}]^2},$$

$$\hat{\mu} = \left[\sum_{j=1}^m \hat{V}_j^{-1} \right]^{-1} \left[\sum_{j=1}^m \hat{V}_j^{-1} Y_j \right],$$

$$\hat{\mu}^{(k)} = \left[\sum_{j=1}^m \hat{V}_j^{-1} \right]^{-1} \left[\sum_{j=1}^m \hat{V}_j^{-1} Y_j^{(k)} \right],$$

$$Y_{ij}^{(k)} = \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} / \sum_{l=1}^{n_j} W_{jl}^{(k)},$$

$$Y_j^{(k)} = [Y_{1j}^{(k)}, \dots, Y_{Ij}^{(k)}]',$$

$$\hat{V}_j = 1/20 \sum_{k=1}^{20} [Y_j^{(k)} - Y_j] [Y_j^{(k)} - Y_j]'. \quad (2.7)$$

Then we propose the following estimator of θ_{ij} :

$$\hat{\theta}_{ij}^* = (1 - \hat{a}_{ij}^*) Y_{ij} + \hat{a}_{ij}^* e_i' \hat{\mu}. \quad (2.8)$$

Remark 4: Using argument given in Remark 3, $\hat{a}_{ij}^* \geq 0$. But it is possible that sometimes \hat{a}_{ij}^* may exceed unity. Thus, we consider the following estimator:

$$\hat{\theta}_{ij} = (1 - \hat{a}_{ij}) Y_{ij} + \hat{a}_{ij} e_i' \hat{\mu}, \quad (2.9)$$

where $\hat{a}_{ij} = \min[1, \hat{a}_{ij}^*]$.

In Table 1, we give values of \hat{a}_{ij} for the West non-self-representing area.

3. ESTIMATION OF THE RELATIVE IMPORTANCE

Let $R_{ij} = Y_{ij} / \sum_{i=1}^I Y_{ij}$ be the preliminary estimator of the relative importance $r_{ij} = \theta_{ij} / \sum_{i=1}^I \theta_{ij}$, ($i = 1, \dots, I; j = 1, \dots, m$). Let $R_j = (R_{1j}, \dots, R_{Ij})'$, ($j = 1, \dots, m$). Since $\sum_{i=1}^I R_{ij} = 1$, ($j = 1, \dots, m$), the variance covariance matrix of R_j is singular. Thus, the method described in Section 2 is not directly applicable to this situation. In order to avoid this singularity problem, we delete one item stratum from the expenditure class under consideration. Without any loss of generality, let the I th item stratum be deleted. Then apply the procedure described in Section 2 to obtain the following estimator for r_{ij} , ($i = 1, \dots, I - 1; j = 1, \dots, m$)

$$\hat{r}_{ij}^* = (1 - \hat{d}_{ij}) R_{ij} + \hat{d}_{ij} e_i' \hat{\xi}, \quad (3.1)$$

where

$$\hat{d}_{ij} = \min[1, \hat{d}_{ij}^*],$$

$$\hat{d}_{ij}^* = \frac{e_i' \left[\hat{D}_j - \left[\sum_{j=1}^m \hat{D}_j^{-1} \right]^{-1} \right] e_i}{\frac{1}{20} \sum_{k=1}^{20} [R_{ij}^{(k)} - e_i' \hat{\xi}^{(k)}]^2},$$

$$R_{ij}^{(k)} = Y_{ij}^{(k)} / \sum_{i=1}^I Y_{ij}^{(k)},$$

$$R_j^{(k)} = (R_{1j}^{(k)}, \dots, R_{I-1j}^{(k)})',$$

$$\hat{D}_j = \frac{1}{20} \sum_{k=1}^{20} (R_j^{(k)} - R_j)(R_j^{(k)} - R_j)',$$

$$\hat{\xi}^{(k)} = \left[\sum_{j=1}^m \hat{D}_j^{-1} \right]^{-1} \left[\sum_{j=1}^m \hat{D}_j^{-1} R_j^{(k)} \right],$$

$$\hat{\xi} = \left[\sum_{j=1}^m \hat{D}_j^{-1} \right]^{-1} \left[\sum_{j=1}^m \hat{D}_j^{-1} R_j \right].$$

For $i = I$,

$$\hat{D}_{II}^{(j)} = \frac{1}{20} \sum_{k=1}^{20} (\hat{R}_{Ij}^{(k)} - R_{Ij})^2,$$

$$R_{I.} = \left[\sum_{j=1}^m (\hat{D}_{II}^{(j)})^{-1} R_{Ij} \right] / \sum_{j=1}^m (\hat{D}_{II}^{(j)})^{-1},$$

$$\hat{d}_{Ij} = \min[1, \hat{d}_{Ij}^*],$$

$$\hat{d}_{Ij}^* = \frac{\hat{D}_{II}^{(j)} - \left[\sum_{j=1}^m (\hat{D}_{II}^{(j)})^{-1} \right]^{-1}}{\frac{1}{20} \sum_{k=1}^{20} [R_{Ij}^{(k)} - R_{I.}^{(k)}]^2},$$

$$R_{I.}^{(k)} = \left[\sum_{j=1}^m (\hat{D}_{II}^{(j)})^{-1} R_{Ij}^{(k)} \right] / \sum_{j=1}^m (\hat{D}_{II}^{(j)})^{-1}.$$

We estimate r_{Ij} by a univariate procedure which yields the following estimator of r_{Ij} , ($j = 1, \dots, m$):

$$\hat{r}_{Ij}^* = (1 - \hat{d}_{Ij})R_{Ij} + \hat{d}_{Ij}R_I.$$

We obtain the final estimator of r_j as $\hat{r}_j = (\hat{r}_{1j}, \dots, \hat{r}_{Ij})'$, where $\hat{r}_{ij} = \hat{r}_{ij}^* / \sum_{i=1}^I \hat{r}_{ij}^*$. There are I possible choices of deleting one item stratum. We choose the combination which yields the smallest average (over item strata) estimated MSE. One may obtain an alternative estimator of r_{Ij} by subtracting $\sum_{i=1}^{I-1} r_{ij}$ from unity. However, according to the procedure, there is a positive probability that r_{Ij} estimate is negative.

4. NUMERICAL RESULTS

In this section, we evaluate various estimators of the cost weight and relative importance based on estimated mean squared error. We consider four rival estimators: the preliminary estimator, estimator proposed by Cohen and Sommers (1984), the estimator currently used by the BLS and the empirical Bayes estimator considered recently by Ghosh and Sohn (1990). The Cohen-Sommers estimator of the cost weight (before raking) is given by

$$\begin{aligned} \hat{\theta}_{ij}^{CS} &= \hat{\theta}_{ij}^{CS*} \quad \text{if} \quad |\hat{\theta}_{ij}^{CS*} - Y_{ij}| < c \cdot \text{sd}(Y_{ij}) \\ &= Y_{ij} + c \cdot \text{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}_{ij}^{CS*} \geq Y_{ij} + c \cdot \text{sd}(Y_{ij}) \\ &= Y_{ij} - c \cdot \text{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}_{ij}^{CS*} \leq Y_{ij} - c \cdot \text{sd}(Y_{ij}) \end{aligned}$$

where

$$\begin{aligned} \hat{\theta}_{ij}^{CS*} &= (1 - \hat{a}_{ij}^{CS})Y_{ij} + \hat{a}_{ij}^{CS}Y_{i.}, \\ \hat{a}_{ij}^{CS} &= \min \left[1, (1 - N_j/N) \left[\frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2 \right] \bigg/ \left[\frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{i.}^{(k)})^2 \right] \right], \\ Y_{i.}^{(k)} &= \sum_{j=1}^m \sum_{l=1}^{n_j} w_{jl}^{(k)} X_{ijl} \bigg/ \sum_{j=1}^m \sum_{l=1}^{n_j} w_{jl}^{(k)}, \end{aligned}$$

N_j = total number of consumer units in the population for the j th publication area,

$$N = \sum_{j=1}^m N_j,$$

$$\text{sd}(Y_{ij}) = \sqrt{\left\{ \frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2 \right\}},$$

c = a safety factor determined by the BLS (see Table 2).

Table 2
Values of the Safety Factor c for the Major Areas

Major Area	NCNS	NCSR	NENS	NESR	SSNS	SSSR	WWNS	WWSR
	1	2	3	4	5	6	7	8
c	1.0	.5	1.0	.5	3.0	.25	1.0	.5

NCNS: North Central (Midwest) non-self-representing.
 NCSR: North Central self-representing.
 NENS: North East non-self-representing.
 SSNS: South non-self-representing.
 SSSR: South self-representing.
 WWNS: West non-self-representing.
 WWSR: West self-representing.

Their estimator for the relative importance is given by

$$\begin{aligned}\hat{r}_{ij}^{CS} &= \hat{r}_{ij}^{CS*} \quad \text{if} \quad |\hat{r}_{ij}^{CS*} - R_{ij}| \leq c \cdot \text{sd}(R_{ij}) \\ &= R_{ij} + c \cdot \text{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{CS*} \geq R_{ij} + c \cdot \text{sd}(R_{ij}) \\ &= R_{ij} - c \cdot \text{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{CS*} \leq R_{ij} - c \cdot \text{sd}(R_{ij}),\end{aligned}$$

where

$$\begin{aligned}\hat{r}_{ij}^{CS*} &= (1 - \hat{d}_{ij}^{CS})R_{ij} + \hat{d}_{ij}^{CS}R_{i.}^{CS}, \\ R_{i.}^{CS} &= \sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl} X_{ijl} / \sum_{i=1}^I \sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl} X_{ijl}, \\ \hat{d}_{ij}^{CS} &= \hat{d}_{ij}^{CS*} \quad \text{if} \quad 0 < \hat{d}_{ij}^{CS*} < 1, \\ &= 0 \quad \text{if} \quad \hat{d}_{ij}^{CS*} \leq 0, \\ &= 1 \quad \text{if} \quad \hat{d}_{ij}^{CS*} \geq 1, \\ \hat{d}_{ij}^{CS*} &= \frac{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2 - \frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})(R_{i.}^{CS(k)} - R_{i.}^{CS})}{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{i.}^{CS(k)})^2}, \\ R_{i.}^{CS(k)} &= \sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} / \sum_{i=1}^I \sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl},\end{aligned}$$

$$sd(R_{ij}) = \sqrt{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2}.$$

Since $\sum_{i=1}^I \hat{r}_{ij}^{CS} \neq 1$, for our comparison purpose, we have divided \hat{r}_{ij}^{CS} by $\sum_{i=1}^I \hat{r}_{ij}^{CS}$.

The procedure currently used by the Bureau of Labor Statistics (see United States Department of Labor 1988) consists of a number of steps.

Step 1: Obtain an estimator of the cost weight as follows:

$$\hat{\theta}_{ij}^{CS(1)} = \hat{r}_{ij}^{CS} \sum_{i=1}^I Y_{ij}.$$

Step 2: Final estimator of θ_{ij} is obtained from $\hat{\theta}_{ij}^{CS(1)}$ using a “raking” procedure. The final estimator, denoted by $\hat{\theta}_{ij}^{BLS}$, satisfies the following two conditions:

$$\begin{aligned} \sum_{i=1}^I \hat{\theta}_{ij}^{BLS} &= \sum_{i=1}^I Y_{ij}, \\ \sum_{j=1}^m N_j \hat{\theta}_{ij}^{BLS} &= \sum_{j=1}^m N_j Y_{ij}. \end{aligned}$$

Step 3: Finally an estimator for the relative importance is obtained as follows:

$$\hat{r}_{ij}^{BLS} = \hat{\theta}_{ij}^{BLS} \bigg/ \sum_{i=1}^I \hat{\theta}_{ij}^{BLS}.$$

In our numerical work, we have estimated N_j by $\sum_{l=1}^{n_j} W_{jl}$.

The MSE of an estimator e_{ij} of θ_{ij} is given by:

$$\begin{aligned} \text{MSE} &= E(e_{ij} - \theta_{ij})^2 \\ &= E(e_{ij} - Y_{ij})^2 - V(Y_{ij}) + 2 \text{Cov}(e_{ij}, Y_{ij}), \end{aligned}$$

where it is assumed $E(Y_{ij} | \theta_{ij}) = \theta_{ij}$. The above formula is given in Cohen and Sommers (1984). As in the Ghosh and Sohn (1990) we estimate the three terms by the balanced repeated half samples available from the survey. For example,

$$\begin{aligned} E(e_{ij} - Y_{ij})^2 &\doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - Y_{ij}^{(k)})^2, \\ V(Y_{ij}) &\doteq \frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2, \end{aligned}$$

Table 3
Average Estimated MSE's for Different Estimators of θ_{ij}

Major Area	Average Estimated MSE of				
	Y_{ij}	$\hat{\theta}_{ij}^{GS}$	$\hat{\theta}_{ij}^{CS}$	$\hat{\theta}_{ij}^{BLS}$	$\hat{\theta}_{ij}$
NCNS	.020047	.011549 (22)	.009342 (53)	.014885 (25)	.009428 (52)
NCSR	.036620	.024783 (32)	.016017 (56)	.023627 (35)	.016155 (55)
NENS	.018162	.013299 (26)	.007327 (59)	.013046 (28)	.005504 (69)
NESR	.052883	.051100 (3)	.038911 (26)	.045610 (13)	.028958 (45)
SSNS	.021757	.013146 (39)	.009954 (54)	.014415 (33)	.006418 (70)
SSSR	.047500	.028984 (38)	.031743 (33)	.044238 (6)	.009270 (80)
WWNS	.052387	.029938 (42)	.017433 (66)	.030069 (42)	.010849 (79)
WWSR	.018223	.033529 (- 83)	.009925 (45)	.014898 (18)	.005761 (68)

Note: The figures in the parenthesis represents percent improvement over the preliminary estimator, Y_{ij} .

$$\text{Cov}(e_{ij}, Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - e_{ij}) (Y_{ij}^{(k)} - Y_{ij}).$$

In the above $e_{ij}^{(k)}$ is the estimator e_{ij} based on the k th half sample ($k = 1, \dots, 20$). For example,

$$\begin{aligned} \hat{\theta}_{ij}^{CS(k)} &= (1 - \hat{a}_{ij}^{CS}) Y_{ij}^{(k)} + \hat{a}_{ij}^{CS} Y_i^{(k)}, \\ \hat{\theta}_{ij}^{(k)} &= (1 - \hat{a}_{ij}) Y_{ij}^{(k)} + \hat{a}_{ij} e_i' \hat{\mu}^{(k)}. \end{aligned}$$

We obtain $\hat{\theta}_{ij}^{BLS(k)}$ by the multistep procedure used to obtain $\hat{\theta}_{ij}^{BLS}$ where we replace Y_{ij} , R_{ij} , \hat{r}_{ij}^{CS} by $Y_{ij}^{(k)}$, $R_{ij}^{(k)}$ and $\hat{r}_{ij}^{CS(k)}$ respectively. Note that the above procedure does not take into account the variation due to the estimation of the coefficients (*i.e.*, a_{ij} 's) in the composite estimators. Cohen and Sommers (1984) recommended the use of half samples of half samples, or quarter samples to capture this additional variability. We could not use their procedure since our dataset did not contain these quarter samples.

The data we analyze arise out of 1982-83 Consumer Expenditure Survey (Diary survey). The expenditure class we consider is dairy products. There are in all six item strata in this class. They are (1) fresh whole milk, (2) other fresh milk and cream, (3) butter, (4) cheese, (5) ice cream and related products, and (6) other dairy products.

The MSE's of all the estimators considered are estimated for each publication area and item stratum. In Table 3 we report the average estimated MSE's of the estimators of θ_{ij} , the average being taken over all the item strata and all the publication areas within a major area. Notice that all the composite estimators except the one proposed by Ghosh and Sohn (1990) are better than the preliminary estimator for all the major areas in the average MSE sense. Both θ_{ij}^{CS} and $\hat{\theta}_{ij}$ are better than $\hat{\theta}_{ij}^{BLS}$. Our proposed estimator $\hat{\theta}_{ij}$ is better than $\hat{\theta}_{ij}^{CS}$ in six out of eight major areas. In two major areas (NCNS and NCSR), $\hat{\theta}_{ij}^{CS}$ is better than $\hat{\theta}_{ij}$, but the difference is very negligible.

In Tables 4 and 5, we try to demonstrate that the raking procedure may not be necessary. In Table 4, the parameter of interest is $\sum_{i=1}^I \theta_{ij}$, the true cost weight for the expenditure class. Here, due to the "raking" procedure, $\sum_{i=1}^I \hat{\theta}_{ij}^{BLS} = \sum_{i=1}^I Y_{ij}$. We propose an alternative estimator as $\sum_{i=1}^I \hat{\theta}_{ij}$ and compare the average estimated MSE (over publication areas in a major area) with that of $\sum_{i=1}^I Y_{ij}$. In all the cases, we gain considerably.

Table 4
Average Estimated MSE's of Two Estimators of Average Consumer
Expenditure for the Expenditure Class

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.12384	0.07969	36
NCSR	0.29819	0.13040	56
NENS	0.21658	0.07602	65
NESR	0.67486	0.20119	70
SSNS	0.21506	0.08303	61
SSSR	0.68415	0.06462	90
WWNS	0.35446	0.05175	85
WWSR	0.19292	0.05524	71

Table 5
Average Estimated MSE's of Two Estimators of Average Consumer
Expenditure for the Major Area

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.008181	0.0045468	44
NCSR	0.003672	0.0031047	15
NENS	0.006174	0.0029128	53
NESR	0.011680	0.0056922	51
SSNS	0.007501	0.0036401	51
SSSR	0.004434	0.0013751	69
WWNS	0.008203	0.0022560	72
WWSR	0.002786	0.0007882	72

In Table 5, the parameter of interest is the cost weight of an item stratum for the major area. The preliminary estimator (identical to the BLS estimator due to the raking procedure) is $(\sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl} Y_{ij}) / (\sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl})$. Our estimation procedure can also generate estimators at the major area level. We propose the estimator as $\hat{\theta}_i = \sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl} \hat{\theta}_{ij} / (\sum_{j=1}^m \sum_{l=1}^{n_j} W_{jl})$. The average estimated MSE's for these two estimators are reported in Table 5. Here also our estimator is superior to the preliminary (BLS) estimator.

The results of Table 4 and 5 suggest that the data consistency step followed by the BLS may not be necessary. Indeed, it may be possible to improve the traditional estimators at higher levels of aggregation also.

Table 6 provides the average estimated MSE's (over all the item strata and publication areas in a major area) of various estimators of relative importance. Notice that as in Table 3, all the estimators other than \hat{r}_{ij}^{GS} are better than the preliminary estimator \hat{R}_{ij} for all the major areas. Our proposed estimator \hat{r}_{ij} is the best among all the estimators considered.

Recently, Swanson (1992) has compared different methods of estimating cost weights for 12 of the approximately 70 expenditure classes in the CPI. His investigation shows that overall our proposed method is superior to all the rival methods.

Table 6
Average Estimated MSE's for Different Estimators of Relative Importance

Major Area	Average Estimated MSE of				
	R_{ij}	\hat{r}_{ij}^{GS}	\hat{r}_{ij}^{CS}	\hat{r}_{ij}^{BLS}	\hat{r}_{ij}
NCNS	.0006342	.00046480 (27)	.00033143 (48)	.00042130 (34)	.00018592 (71)
NCSR	.0009125	.00071967 (21)	.00040226 (56)	.00044815 (51)	.00021309 (77)
NENS	.0003588	.00026894 (25)	.00014146 (61)	.0001620 (55)	.00011105 (69)
NESR	.0004264	.00072001 (- 69)	.00028862 (32)	.00030555 (28)	.00016744 (61)
SSNS	.0005071	.00033736 (33)	.00019352 (62)	.00021385 (58)	.00011925 (76)
SSSR	.0006564	.00048569 (26)	.00053173 (19)	.00053603 (18)	.00030979 (53)
WWNS	.0013709	.00086849 (37)	.00051474 (62)	.00061901 (55)	.00028519 (79)
WWSR	.0003540	.00070770 (- 100)	.00021384 (40)	.00023255 (34)	.00013750 (61)

Note: The figure given in the parenthesis represents percent improvement over R_{ij} .

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation (NSF) under grant SES-9001399, "On-Site Research to Improve the Quality of Labor Statistics." This research was conducted at the U.S. Bureau of Labor Statistics while the authors were participants in the American Statistical Association/Bureau of Labor Statistics Research Program, which is supported by the Bureau of Labor Statistics and through the NSF grant. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of Labor Statistics. We wish to thank Sylvia Leaver, Stuart Scott, Malay Ghosh, Richard Valliant, Stephen Miller, So Young Sohn, Paul Hsen, Adriana Silberstein for many valuable discussions. We also thank two referees and associate editor for helpful comments on an earlier version of this paper.

REFERENCES

- COHEN, M.P., and SOMMERS, J.P. (1984). Evaluation of methods of composite estimation of cost weights for the CPI. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 466-471.
- GHOSH, M. and SOHN, S.Y. (1990). An Empirical Bayes Approach Towards Composite Estimation of Consumer Expenditure. Technical Report, U.S. Bureau of Labor Statistics.
- MARKS, H. (1978). Composite estimation techniques used for the CPIR weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 311-315.
- MCCARTHY P.J. (1969). Pseudoreplication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.
- SWANSON, D. (1992). An evaluation of 4 cost weight composite estimation methods for the CPI. Memorandum for Janet Williams, Chief, CPI Survey Research Branch, Statistical Methods Division, U.S. Bureau of Labor Statistics.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd Edition). New York: J. Wiley & Sons.
- UNITED STATES DEPARTMENT OF LABOR (1988). *Handbook of Methods*. Bureau of Labour Statistics. Washington DC: U.S. Government Printing Office.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1992:

- P. Ardilly, *INSEE*
M.G. Arellano, *Advanced Linkage Technologies of America*
K.G. Basavarajappa, *Statistics Canada*
T.R. Belin, *UCLA*
D.R. Bellhouse, *University of Western Ontario*
P. Biemer, *Research Triangle Institute*
D. Binder, *Statistics Canada*
P.D. Bourke, *University College, Cork, Ireland*
J.M. Brick, *Westat*
N.J. Carter, *California State University*
R.G. Carter, *Statistics Canada*
G.H. Choudhry, *Statistics Canada*
P.A. Cholette, *Statistics Canada*
M.P. Cohen, *U.S. National Center for Education Statistics*
B. Cox, *Research Triangle Institute*
E.B. Dagum, *Statistics Canada*
J.-C. Deville, *INSEE*
C. Dippo, *U.S. Bureau of Labor Statistics*
D. Drew, *Statistics Canada*
R.E. Fay, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
M. Frankel, *Baruch College, CUNY*
J. Gentleman, *Statistics Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R. Groves, *U.S. Bureau of the Census*
K.P. Hapuarachchi, *Statistics Canada*
M.A. Hidirolglou, *Statistics Canada*
D. Holt, *University of Southampton*
P. Jagers, *Chalmers and Gothenburg Universities*
G. Kalton, *University of Michigan*
P.S. Kott, *NASS/U.S. Department of Agriculture*
R.A. Kulka, *University of Chicago*
P. Lahiri, *University of Nebraska*
G. Lagrange, *Statistics Canada*
K.C. Land, *Duke University*
J.M. Landwehr, *AT & T Bell Laboratories*
N. Laniel, *Statistics Canada*
P. Lavallée, *Statistics Canada*
H. Lee, *Statistics Canada*
F. Maranda, *Statistics Canada*
M. March, *Statistics Canada*
G.D. Meeden, *University of Minnesota*
W.J. Mitofsky, *Voter Research and Surveys*
H.B. Newcombe, *Consultant*
W.L. Nicholls II, *U.S. Bureau of the Census*
D. Norris, *Statistics Canada*
D. Northrup, *Coda Inc.*
C. O'Muircheartaigh, *London School of Economics and Political Science*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistics Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
D. Royce, *Statistics Canada*
D. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
N.C. Schaiffer, *University of Wisconsin - Madison*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York - Albany*
G.M. Shapiro, *U.S. Bureau of the Census*
C.J. Skinner, *University of Southampton*
B.D. Spencer, *Northwestern University*
N.L. Spruill, *U.S. Office of the Secretary of Defence*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
H. Tamura, *University of Washington*
A. Thériège, *Statistics Canada*
M.E. Thompson, *University of Waterloo*
R. Valliant, *U.S. Bureau of Labor Statistics*
K. Wachter, *University of California - Berkeley*
J. Waksberg, *Westat*
T. Wellens, *Zentrum für Umfragen, Zuma*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*

Acknowledgements are also due to those who assisted during the production of the 1992 issues: J. Beauseigle, S. Beauchamp and S. Lineger (Photocomposition), and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

**Special Issue on Measurement Errors in Surveys: Part I
Contents JOS 1992, Volume 8, Number 1**

Preface	3
Cognitive Aspects of Surveys: Yesterday, Today, and Tomorrow <i>Judith M. Tanur and Stephen E. Fienberg</i>	5
Measuring the Recall Error in Self-Reported Fishing and Hunting Activities <i>Adam Chu, Donna Eisenhower, Michael Hay, David Morganstein, John Neter, and Joseph Waksberg</i>	19
The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Diary Survey <i>Clyde Tucker</i>	41
Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility <i>Mick P. Couper, Lisa Holland, and Robert M. Groves</i>	63
The Golden Numerical Comparative Scale Format for Economical Multi-Object/Multi-Attribute Comparison Questionnaires <i>Linda L. Golden, Patrick L. Brackett, Gerald Albaum, and Juan Zatarain</i>	77
Effects of Procedural Differences in the Nationwide Food Consumption Survey <i>Patricia M. Guenther</i>	87
Evidence of Anchoring in a Survey Recall Task <i>Carolyn M. Boyce and Marilyn C. Mauch</i>	97
Special Notes	105
In Other Journals	107
Book Reviews	109

All inquiries about submissions and subscriptions should be directed to the Chief Editor:

Lars Lyberg, U/SFI, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 41, No. 3, 1992

	<i>Page</i>
Statistical inference in crime investigations using deoxyribonucleic acid profiling (with discussion) <i>D. A. Berry, I. W. Evett and R. Pinchin</i>	499
Ranking methods for compositional data <i>J. Bacon-Shone</i>	533
Assessing the nature of periodontal disease progression—an application of covariance structure estimation <i>J. A. C. Sterne, A. Kingman and H. Loe</i>	539
Box–Cox transformations and the Taguchi method: an alternative analysis of a Taguchi case study <i>T. Fearn</i>	553
<i>General Interest Section</i>	
Subjective modelling and Bayes linear estimation in the UK water industry <i>A. O'Hagan, E. B. Glennie and R. E. Beardsall</i>	563
Modelling variation in industrial experiments <i>J. Engel</i>	579
<i>Letters to the Editors</i>	595
<i>Book Reviews</i>	601
<i>Statistical Software Review</i>	
SOLO	605
<i>Statistical Algorithms</i>	
AS 277 The Oja bivariate median <i>A. Niinimaa, H. Oja and J. Nyblom</i>	611
AS 278 Distribution of quadratic forms of multivariate generalized Student variables <i>B. Lecoutre, J.-L. Guigues and J. Poitevineau</i>	617
<i>Remark</i>	
AS R90 Least squares initial values for the L_1 -norm fitting of a straight line—a remark on Algorithm AS 238: A simple recursive procedure for the L_1 norm fitting of a straight line <i>R. W. Farebrother</i>	627
<i>Correction</i>	
Correction to Algorithm AS 271: General optimal combinatoric classification <i>C. L. Dunn</i>	634
<i>Author Index</i>	635

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1.

Présentation

1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3

Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4

Les remerciements doivent paraître à la fin du texte.

1.5

Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2.

Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.

Rédaction

3.1

Éviter les notes au bas des pages, les abréviations et les sigles.

3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5

Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, I).

3.6

Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4.

Figures et tableaux

4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5.

Bibliographie

5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.

5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1992:

P. Ardilly, *INSEE*
M.G. Arellano, *Advanced Linkage Technologies of America*
K.G. Basavarajappa, *Statistique Canada*
T.R. Belin, *UCLA*
D.R. Bellhouse, *University of Western Ontario*
P. Biemer, *Research Triangle Institute*
D. Binder, *Statistique Canada*
P.D. Bourke, *University College, Cork, Ireland*
J.M. Brick, *Westat*
N.J. Carter, *California State University*
R.G. Carter, *Statistique Canada*
G.H. Choudhry, *Statistique Canada*
P.A. Cholette, *Statistique Canada*
M.P. Cohen, *U.S. National Center for Education Statistics*
B. Cox, *Research Triangle Institute*
E.B. Dagum, *Statistique Canada*
J.-C. Deville, *INSEE*
C. Diplo, *U.S. Bureau of Labor Statistics*
D. Drew, *Statistique Canada*
R.E. Fay, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
M. Frankel, *Baruch College, CUNY*
J. Gentleman, *Statistique Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R. Groves, *U.S. Bureau of the Census*
K.P. Hapuarachchi, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *University of Southampton*
P. Jagers, *Chalmers and Gothenburg Universities*
G. Kalton, *University of Michigan*
P.S. Kott, *NASS/U.S. Department of Agriculture*
R.A. Kulka, *University of Chicago*
P. Lahiri, *University of Nebraska*
G. Lagrange, *Statistique Canada*
K.C. Land, *Duke University*
J.M. Landwehr, *AT & T Bell Laboratories*
N. Laniel, *Statistique Canada*
P. Lavallée, *Statistique Canada*

H. Lee, *Statistique Canada*
F. Maranda, *Statistique Canada*
M. March, *Statistique Canada*
G.D. Meeden, *University of Minnesota*
W.J. Mitofsky, *Voter Research and Surveys*
H.B. Newcombe, *Expert Conseil*
W.L. Nichols II, *U.S. Bureau of the Census*
D. Norris, *Statistique Canada*
D. Northrup, *Coda Inc.*
C. O'Muircheartaigh, *London School of Economics and Political Science*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistique Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
G. Roberts, *Statistique Canada*
D. Royce, *Statistique Canada*
D. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
N.C. Schaffter, *University of Wisconsin - Madison*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York - Albany*
G.M. Shapiro, *U.S. Bureau of the Census*
C.J. Skinner, *University of Southampton*
B.D. Spencer, *Northwestern University*
N.L. Sprull, *U.S. Office of the Secretary of Defence*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
H. Tamura, *University of Washington*
A. Thèberge, *Statistique Canada*
M.E. Thompson, *University of Waterloo*
R. Valliant, *U.S. Bureau of Labor Statistics*
K. Wachter, *University of California - Berkeley*
J. Waksberg, *Westat*
T. Welles, *Zentrum für Umfragen, Zuma*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1992: J. Beauséjle, S. Beauchamp et S. Liniger (Photocomposition), et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à S. Di Loreto, M. Kent, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

BIBLIOGRAPHIE

COHEN, M.P., et SOMMERS, J.P. (1984). Evaluation of methods of composite estimation of cost weights for the CPI. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 466-471.

GHOSH, M. et SOHN, S.Y. (1990). An Empirical Bayes Approach Towards Composite Estimation of Consumer Expenditure. Rapport technique, U.S. Bureau of Labor Statistics.

MARKS, H. (1978). Composite estimation techniques used for the CPIR weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 311-315.

McCARTHY P.J. (1969). Pseudoreplication: half-samples. *Revue de l'Institut International de Statistique*, 37, 239-264.

SWANSON, D. (1992). An evaluation of 4 cost weight composite estimation methods for the CPI. Note de service pour Janet Williams, Chief, CPI Survey Research Branch, Statistical Methods Division, U.S. Bureau of Labor Statistics.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2^{ème} Edition). New York: J. Wiley & Sons.

UNITED STATES DEPARTMENT OF LABOR (1988). *Handbook of Methods*. Bureau of Labour Statistics. Washington DC: U.S. Government Printing Office.

Tableau 6

Erreur quadratique moyenne estimé de l'importance relative				
Région Principale	Erreur quadratique moyenne estimé de			
	R_{ij}	f_{ij}^{GS}	f_{ij}^{CS}	f_{ij}^{BLS}
				f_{ij}

NCNS	.0006342	.00046480	.00033143	.00042130	.00018592
	(27)	(48)	(34)	(71)	
NCSR	.0009125	.00071967	.00040226	.00044815	.00021309
	(21)	(56)	(51)	(77)	
NENS	.0003588	.00026894	.00014146	.0001620	.00011105
	(25)	(61)	(55)	(69)	
NESR	.0004264	.00072001	.00028862	.00030555	.00016744
	(-69)	(32)	(28)	(61)	
SSNS	.0005071	.00033736	.00019352	.00021385	.00011925
	(33)	(62)	(58)	(76)	
SSSR	.0006564	.00048569	.00053173	.00053603	.00030979
	(26)	(19)	(18)	(53)	
WWNS	.0013709	.00086849	.00051474	.00061901	.00028519
	(37)	(62)	(55)	(79)	
WWSR	.0003540	.00070770	.00021384	.00023255	.00013750
	(-100)	(40)	(34)	(61)	

Nota: Les chiffres entre parenthèses représentent l'amélioration en pourcentage par rapport à R_{ij} .

REMERCIEMENTS

Cet article est le résultat de recherches qui ont été rendues possibles grâce à une subvention de la National Science Foundation (NSF) (SES-9001399 – “On-Site Research to Improve the Quality of Labor Statistics.”) Ces recherches ont été faites au Bureau of Labor Statistics des E.-U. pendant que les auteurs participaient au programme conjoint de recherches de l’American Statistical Association et du Bureau of Labor Statistics, financé par le BLS et la NSF. Les opinions, conclusions ou recommandations qui peuvent être exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position de la National Science Foundation ni celle du Bureau of Labor Statistics. Nous tenons à remercier Sylvia Leaver, Stuart Scott, Malay Ghosh, Richard Valliant, Stephen Miller, So Young Sohn, Paul Hsen et Adriana Silberstein pour les nombreux et précieux échanges de vues que nous avons eus avec ces personnes. Nous voulons exprimer aussi notre reconnaissance aux deux arbitres et au rédacteur associé qui ont formulé des commentaires utiles sur une version antérieure de cet article.

Tableau 4
 Erreur quadratique moyenne estimée de deux estimateurs des dépenses
 de consommation moyennes pour une catégorie de dépenses

Région principale	Estimateur préliminaire	Estimateur proposé	Amélioration (en pourcentage)
NCNS	0.12384	0.07969	36
NCSR	0.29819	0.13040	56
NENS	0.21658	0.07602	65
NESR	0.67486	0.20119	70
SSNS	0.21506	0.08303	61
SSSR	0.68415	0.06462	90
WNNS	0.35446	0.05175	85
WSR	0.19292	0.05524	71

Tableau 5
 Erreur quadratique moyenne estimée de deux estimateurs des dépenses
 de consommation moyennes pour une région principale

Région principale	Estimateur préliminaire	Estimateur proposé	Amélioration (en pourcentage)
NCNS	0.008181	0.0045468	44
NCSR	0.003672	0.0031047	15
NENS	0.006174	0.0029128	53
NESR	0.011680	0.0056922	51
SSNS	0.007501	0.0036401	51
SSSR	0.004434	0.0013751	69
WNNS	0.008203	0.0022560	72
WSR	0.002786	0.0007882	72

Le tableau 6 donne l'erreur quadratique moyenne estimée (pour l'ensemble des strates et l'ensemble des régions de publication d'une région principale) de divers estimateurs de l'importance relative. Notons que, comme dans le tableau 3, tous les estimateurs sauf f_{ij}^{GS} surclassent l'estimateur préliminaire R_{ij} pour toutes les régions principales. L'estimateur que nous proposons, f_{ij} , est le plus efficace de tous les estimateurs étudiés.

Dernièrement, Swanson (1992) a comparé différentes méthodes d'estimation du poids en valeur pour 12 des quelque 70 catégories de dépenses de l'IPC. Son étude montre que dans l'ensemble, la méthode que nous proposons est supérieure à toutes les autres.

Tableau 3

Erreur quadratique moyenne estimée de différents estimateurs de θ_{ij}

Région principale	Erreur quadratique moyenne de			
	Y_{ij}	θ_{GS}^{ij}	θ_{CS}^{ij}	θ_{BLS}^{ij}
NCNS	.020047	.011549	.009342	.014885
NCNS	(22)	(53)	(25)	(52)
NCSR	.036620	.024783	.016017	.023627
NCSR	(32)	(56)	(35)	(55)
NENS	.018162	.013299	.007327	.013046
NENS	(26)	(59)	(28)	(69)
NESR	.052883	.051100	.038911	.045610
NESR	(3)	(26)	(13)	(45)
SSNS	.021757	.013146	.009954	.014415
SSNS	(39)	(54)	(33)	(70)
SSSR	.047500	.028984	.031743	.044238
SSSR	(38)	(33)	(6)	(80)
WVNS	.052387	.029938	.017433	.030069
WVNS	(42)	(66)	(42)	(79)
WWSR	.018223	.033529	.009925	.014898
WWSR	(-83)	(45)	(18)	(68)

Nota: Les chiffres entre parenthèses représentent l'amélioration en pourcentage par rapport à l'estimateur préliminaire, Y_{ij} .

supérieurs à θ_{BLS}^{ij} . L'estimateur que nous proposons, θ_{ij} , surclasse θ_{CS}^{ij} dans six des huit régions principales. Dans les deux autres régions (NCNS et NCSR), θ_{ij} est supérieur, mais de très peu, à θ_{ij} .

Dans les tableaux 4 et 5, nous tentons de démontrer que le procédé de "balayage" n'est peut-être pas nécessaire. Le tableau 4 a pour objet le paramètre $\sum_{i=1}^I \theta_{ij}$, c.-à-d. le poids en valeur vrai pour une catégorie de dépenses. À cause du procédé de "balayage", $\sum_{i=1}^I \theta_{BLS}^{ij} = \sum_{i=1}^I Y_{ij}$. Voici que nous proposons un autre estimateur, $\sum_{i=1}^I \theta_{ij}$, et que nous comparons son erreur quadratique moyenne estimée (pour l'ensemble des régions de publication d'une région principale) à celle de $\sum_{i=1}^I Y_{ij}$. Dans tous les cas, notre estimateur est notablement supérieur.

Le tableau 5 a pour objet le poids en valeur d'une strate de produits et services pour une région principale. L'estimateur préliminaire (qui est identique à l'estimateur BLS en raison du procédé de "balayage"), s'écrit $(\sum_{j=1}^J W_{ij}^{ij} Y_{ij}) / ((\sum_{j=1}^J W_{ij}^{ij} W_{ij}^{ij}))$. La méthode que nous présentons permet aussi la construction d'estimateurs au niveau de la région principale. L'estimateur que nous proposons est $\theta_{ij} = \sum_{j=1}^J W_{ij}^{ij} W_{ij}^{ij} / ((\sum_{j=1}^J W_{ij}^{ij} W_{ij}^{ij}))$. On constate une fois de plus que notre estimateur surclasse l'estimateur préliminaire (BLS).

Les résultats des tableaux 4 et 5 donnent à penser que l'étape de l'harmonisation des données par un procédé de "balayage", comme le prévoit la méthode du BLS, n'est peut-être pas indispensable. En effet, on peut vraisemblablement améliorer les estimateurs classiques à des niveaux d'aggrégation supérieurs.

Dans nos calculs, nous avons estimé N_j par $\sum_{i=1}^{n_j} W_{ji}$.
L'erreur quadratique moyenne d'un estimateur e_{ij} de θ_{ij} est définie par l'expression

$$EQM = E(e_{ij} - \theta_{ij})^2$$

$$= E(e_{ij} - Y_{ij})^2 - V(Y_{ij}) + 2 \text{Cov}(e_{ij}, Y_{ij}),$$

où l'on suppose que $E(Y_{ij} | \theta_{ij}) = \theta_{ij}$. La formule ci-dessus est définie dans Cohen et Sommers (1984). Comme dans Ghosh et Sohn (1990), nous estimons les trois termes de la formule au moyen des demi-échantillons compensés de l'enquête. Par exemple,

$$E(e_{ij} - Y_{ij})^2 \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - Y_{ij}^{(k)})^2,$$

$$V(Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2,$$

$$\text{Cov}(e_{ij}, Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - e_{ij})(Y_{ij}^{(k)} - Y_{ij}).$$

Dans l'expression ci-dessus, $e_{ij}^{(k)}$ désigne l'estimateur e_{ij} fondé sur le k -ième demi-échantillon ($k = 1, \dots, 20$). Par exemple,

$$\theta_{CS}^{ij(k)} = (1 - a_{ij}^{ij}) Y_{ij}^{(k)} + a_{ij}^{ij} Y_{CS}^{(k)},$$

$$\theta_{ij}^{(k)} = (1 - a_{ij}) Y_{ij}^{(k)} + a_{ij} e_{ij}^{(k)}.$$

Nous calculons $\theta_{BLS}^{ij(k)}$ à l'aide de la méthode multi-étape utilisée pour calculer θ_{BLS}^{ij} , sauf que nous remplaçons $Y_{ij}^{(k)}$ et $f_{CS}^{ij(k)}$ par $Y_{ij}^{(k)}$ et $f_{CS}^{ij(k)}$ respectivement. Il convient de souligner que la procédure ci-dessus ne tient pas compte de la variation due à l'estimation des coefficients (c.-à-d. des a_{ij}) des estimateurs composites. Cohen et Sommers (1984) ont recommandé l'utilisation de demi-échantillons des demi-échantillons, ou quarts d'échantillons, pour tenir compte de cette variation. Nous n'avons pas pu appliquer cette méthode puisque notre ensemble de données ne renfermait pas de données relatives à des quarts d'échantillons.

Les données que nous analysons sont tirées de la Consumer Expenditure Survey de 1982-1983 (volet "enquête journal"). La catégorie de dépenses étudiée est celle des produits laitiers. Cette catégorie compte en tout six strates de produits et services: 1) lait entier frais, 2) autres sortes de lait frais et crème, 3) beurre, 4) fromage, 5) crème glacée et produits assimilés et 6) autres produits laitiers.

Nous estimons l'erreur quadratique moyenne de tous les estimateurs étudiés pour chaque région de publication et chaque strate de produits et services. Le tableau 3 donne l'erreur quadratique moyenne estimée (pour l'ensemble des strates et l'ensemble des régions de publication d'une région principale) des estimateurs de θ_{ij} . Notons que sur ce plan, tous les estimateurs composites sauf celui proposé par Ghosh et Sohn (1990) surclassent l'estimateur préliminaire pour toutes les régions principales. Les estimateurs θ_{CS}^{ij} et θ_{ij} sont tous deux

$$d_{ij}^{CS} = d_{ij}^{CS*} \quad \text{si } 0 < d_{ij}^{CS*} < 1,$$

$$= 0 \quad \text{si } d_{ij}^{CS*} \leq 0,$$

$$= 1 \quad \text{si } d_{ij}^{CS*} \geq 1,$$

$$d_{ij}^{CS*} = \frac{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2 - \frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)})(R_{ij}^{CS(k)})}{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)})^2 - R_{ij}^{CS*2}},$$

$$R_{ij}^{CS(k)} = \sum_{m=1}^m \sum_{n_j=1}^{n_j} W_{ijl}^{f(k)} X_{ijl} / \left| \sum_{l=1}^l \sum_{m=1}^m \sum_{n_j=1}^{n_j} W_{ijl}^{f(k)} X_{ijl} \right|,$$

$$sd(R_{ij}) = \sqrt{\frac{1}{20} \sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2}.$$

Puisque $\sum_{l=1}^l f_{ij}^{CS} \neq 1$, nous avons divisé f_{ij}^{CS} par $\sum_{l=1}^l f_{ij}^{CS}$ pour les besoins de la comparaison. La méthode qu'utilise actuellement le Bureau of Labor Statistics (voir United States Department of Labor 1988) comporte un certain nombre d'étapes.

Étape 1 : Déterminer un estimateur du poids en valeur comme ci-dessous:

$$\theta_{ij}^{CS(1)} = f_{ij}^{CS} \sum_{l=1}^l Y_{ij}.$$

Étape 2 : Dédurre l'estimateur final de θ_{ij} de $\theta_{ij}^{CS(1)}$ par un procédé de "balayage". L'estimateur final, désigné par θ_{ij}^{BLS} , satisfait aux deux conditions suivantes:

$$\sum_{l=1}^l \theta_{ij}^{BLS} = \sum_{l=1}^l Y_{ij},$$

$$\sum_{m=1}^m N_j \theta_{ij}^{BLS} = \sum_{m=1}^m N_j Y_{ij}.$$

Étape 3 : Enfin, déterminer un estimateur de l'importance relative comme ci-dessous:

$$f_{ij}^{BLS} = \theta_{ij}^{BLS} / \sum_{l=1}^l \theta_{ij}^{BLS}.$$

Tableau 2

Valeurs du facteur de sécurité (c) pour les régions principales

Région principale	c								
	1	2	3	4	5	3.0	.25	1.0	.5
NCNS									
NCSR									
NENS									
NESR									
SSNS									
SSSR									
WWSR									

NCNS: Centre-Nord (Midwest) - non autoreprésentative.
NCSR: Centre-Nord - autoreprésentative.
NENS: Nord-Est - non autoreprésentative.
NESR: Nord-Est - autoreprésentative.
SSNS: Sud - non autoreprésentative.
SSSR: Sud - autoreprésentative.
WWSR: Ouest - non autoreprésentative.
WWSR: Ouest - autoreprésentative.

$$Y_{ij}^{(k)} = \sum_{m=1}^f \sum_{n_j^f} W_{n_j^f}^f X_{ijl}^f / \sum_{m=1}^f \sum_{n_j^f} W_{n_j^f}^{f(k)},$$

N_j = nombre total d'unités de consommation dans la région de publication j,

$$N = \sum_{m=1}^f N_j,$$

$$sd(Y_{ij}) = \sqrt{\left\{ \frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij}^2) \right\}},$$

c = facteur de sécurité déterminé par le BLS (voir tableau 2).

L'estimateur de Cohen-Sommers de l'importance relative est défini par l'expression

$$f_{ij}^{CS} = f_{ij}^{CS*} \quad \text{si} \quad |f_{ij}^{CS*} - R_{ij}| \leq c \cdot sd(R_{ij})$$

$$= R_{ij} + c \cdot sd(R_{ij}) \quad \text{si} \quad f_{ij}^{CS*} \geq R_{ij} + c \cdot sd(R_{ij})$$

$$= R_{ij} + c \cdot sd(R_{ij}) \quad \text{si} \quad f_{ij}^{CS*} \leq R_{ij} - c \cdot sd(R_{ij}),$$

où

$$f_{ij}^{CS*} = (1 - d_{ij}^{CS}) R_{ij} + d_{ij}^{CS} R_{ij}^{CS},$$

$$R_{ij}^{CS} = \sum_{m=1}^f \sum_{n_j^f} W_{n_j^f}^f X_{ijl}^f / \sum_{m=1}^f \sum_{n_j^f} W_{n_j^f}^f X_{ijl}^f,$$

$$d_{lj} = \min [1, d_{lj}^*],$$

$$d_{lj}^* = \frac{D_{(j)}^H - \left[\sum_m^j (D_{(j)}^H)^{-1} \right]^{-1}}{\frac{1}{20} \sum_{k=1}^{20} [R_{lj}^{(k)} - R_{(k)}^H]^2},$$

$$R_{(k)}^H = \left[\sum_m^j (D_{(j)}^H)^{-1} R_{lj}^{(k)} \right] / \left[\sum_m^j (D_{(j)}^H)^{-1} \right].$$

Pour estimer r_{lj} , nous nous servons d'une méthode univariée qui donne l'estimateur de r_{lj} , suivant, ($j = 1, \dots, m$):

$$r_{lj}^* = (1 - d_{lj})R_{lj} + d_{lj}R_j.$$

Comme estimateur final de r_j nous avons $\hat{r}_j = (\hat{r}_{lj}, \dots, \hat{r}_{lj})'$, où $\hat{r}_{lj} = r_{lj}^* / \sum_{l=1}^L r_{lj}^*$. Pour ce qui est de la suppression de la strate de produits et services, nous avons L choix possibles. Nous choisissons la combinaison pour laquelle l'erreur quadratique moyenne estimée (pour l'ensemble des strates) est la plus faible. On peut obtenir un autre estimateur de r_{lj} en sous-trayant $\sum_{l=1}^L r_{lj}$ de un. Cependant, selon la procédure, il y a des chances réelles que la valeur estimée de r_{lj} soit négative.

4. RÉSULTATS NUMÉRIQUES

Dans cette section, nous évaluons divers estimateurs du poids en valeur et de l'importance relative en nous fondant sur l'erreur quadratique moyenne estimée. Nous comparons en fait quatre estimateurs: l'estimateur préliminaire, l'estimateur proposé par Cohen et Sommers (1984), l'estimateur qu'utilise actuellement le BLS et l'estimateur empirique de Bayes qu'ont étudié récemment Ghosh et Sohn (1990). L'estimateur de Cohen-Sommers du poids en valeur (avant balayage) est défini par l'expression

$$\hat{\theta}_{CS}^{lj} = \hat{\theta}_{CS}^{lj*} \quad \text{si} \quad |\hat{\theta}_{CS}^{lj*} - Y_{lj}| > c \cdot \text{sd}(Y_{lj})$$

$$= Y_{lj} + c \cdot \text{sd}(Y_{lj}) \quad \text{si} \quad \hat{\theta}_{CS}^{lj*} \geq Y_{lj} + c \cdot \text{sd}(Y_{lj})$$

$$= Y_{lj} - c \cdot \text{sd}(Y_{lj}) \quad \text{si} \quad \hat{\theta}_{CS}^{lj*} \leq Y_{lj} - c \cdot \text{sd}(Y_{lj})$$

où

$$\hat{\theta}_{CS}^{lj*} = \min \left[1, \left(1 - \frac{1}{N_{j(N)}} \right) \left[\frac{1}{20} \sum_{k=1}^{20} Y_{lj}^{(k)} - \frac{1}{20} \sum_{k=1}^{20} \frac{Y_{lj}^{(k)}}{\left[\frac{1}{20} \sum_{k=1}^{20} Y_{lj}^{(k)} \right]^2} \right] \right],$$

$\sum_{j=1}^I R_{ij} = 1, (j = 1, \dots, m)$, la matrice de variances-covariances de R_j est singulière. On ne peut donc appliquer directement la méthode exposée dans la section précédente. Pour contourner le problème de la singularité, nous supprimons une strate de produits et services de la catégorie de dépenses étudiée. Sans aucune perte de généralité, supprimons la strate 1. Dans un deuxième temps, appliquons la méthode décrite dans la section 2 afin d'obtenir l'estimateur de r_{ij} suivant, ($i = 1, \dots, I - 1; j = 1, \dots, m$)

(3.1) $r_{ij}^* = (1 - d_{ij})R_{ij} + d_{ij}e_i'\xi,$

où

$$d_{ij} = \min [1, d_{ij}^*],$$
$$d_{ij}^* = \frac{e_i' \left[D_j - \left[\sum_m D_j^{-1} \right]^{-1} \right] e_i}{\sum_{k=1}^{20} \frac{1}{20} [R_{ij(k)}^t - e_i' \xi_{(k)}^t]^2},$$
$$R_{ij(k)}^t = Y_{ij(k)}^t / \sum_I Y_{ij(k)}^t,$$
$$R_{j(k)}^t = (R_{1j(k)}^t, \dots, R_{I-1j(k)}^t)',$$
$$D_j = \frac{1}{20} \sum_{k=1}^{20} (R_{j(k)}^t (R_{j(k)}^t)' - R_j^t)' ,$$
$$\xi_{j(k)}^t = \left[\sum_m D_j^{-1} \right]^{-1} \left[\sum_m D_j^{-1} R_{j(k)}^t \right],$$
$$\xi = \left[\sum_m D_j^{-1} \right]^{-1} \left[\sum_m D_j^{-1} R_j \right].$$

Pour $i = I,$

$$D_{(j)}^H = \frac{1}{20} \sum_{k=1}^{20} (R_{Ij(k)}^H)' (R_{Ij(k)}^H - R_{Ij})',$$
$$R_{Ij} = \left[\sum_m D_{(j)}^H \right]^{-1} \left[\sum_m D_{(j)}^H (D_{(j)}^H)^{-1} R_{Ij} \right] / \sum_m D_{(j)}^H,$$

Nous allons estimer \hat{a}_{ij} à l'aide des 20 demi-échantillons compensés provenant de l'enquête. Posons $w_{ij}^{(k)}$ comme le poids attribué à l'unité de consommation l de la région j pour l'échantillon répété k ($j = 1, \dots, m; l = 1, \dots, n_j; k = 1, \dots, 20$). Ces poids sont élaborés par le Census Bureau selon une procédure complexe. Pour un échantillon répété quelconque, environ la moitié des unités de consommation reçoivent un poids nul et le reste reçoivent un poids positif. Définissons

$$\hat{a}_{ij}^* = \frac{e_i' \left[\mathcal{P}_j - \left[\sum_{m=1}^f \mathcal{P}_j^{-1} \right]^{-1} \right] e_i}{\frac{1}{20} \sum_{k=1}^{20} [Y_{ij}^{(k)} - e_i' \hat{h}_{ij}^{(k)}]^2},$$

$$\hat{h}_{ij} = \left[\sum_{m=1}^f \mathcal{P}_j^{-1} \right]^{-1} \left[\sum_{m=1}^f \mathcal{P}_j^{-1} X_j \right],$$

$$\hat{h}_{ij}^{(k)} = \left[\sum_{m=1}^f \mathcal{P}_j^{-1} \right]^{-1} \left[\sum_{m=1}^f \mathcal{P}_j^{-1} X_j^{(k)} \right],$$

$$X_j^{(k)} = \sum_{n_j^l}^l w_{ij}^{(k)} X_{ijl}^{(k)} / \sum_{n_j^l}^l w_{ij}^{(k)},$$

$$Y_j^{(k)} = [X_j^{(k)}, \dots, X_j^{(k)}]'$$

$$\mathcal{P}_j = 1/20 \sum_{k=1}^{20} [X_j^{(k)} - X_j] [X_j^{(k)} - X_j]'$$

Nous proposons alors l'estimateur de θ_{ij} suivant:

$$(2.8) \quad \hat{\theta}_{ij}^* = (1 - \hat{a}_{ij}^*) Y_{ij} + \hat{a}_{ij}^* e_i' h_{ij}.$$

Remarque 4: Sur la base de l'argument donné dans la remarque 3, $\hat{a}_{ij}^* \geq 0$. Mais il peut arriver que \hat{a}_{ij}^* soit supérieur à un. Par conséquent, nous considérons l'estimateur suivant:

$$(2.9) \quad \hat{\theta}_{ij} = (1 - \hat{a}_{ij}) Y_{ij} + \hat{a}_{ij} e_i' h_{ij},$$

où $\hat{a}_{ij} = \min [1, \hat{a}_{ij}^*]$. Le tableau 1 donne les valeurs de \hat{a}_{ij} pour la région non autoreprésentative de l'Ouest.

3. ESTIMATION DE L'IMPORTANCE RELATIVE

Soit $R_{ij} = Y_{ij} / \sum_{l=1}^l Y_{ijl}$ l'estimateur préliminaire de l'importance relative $r_{ij} = \theta_{ij} / \sum_{l=1}^l \theta_{ijl}$, ($i = 1, \dots, I; j = 1, \dots, m$). Posons $R_j = (R_{1j}, \dots, R_{Ij})'$, ($j = 1, \dots, m$). Comme

Remarque 1 : Dans le calcul de l'estimateur optimal $\hat{\theta}_{ij}$, on suppose que les quantités V_j ($j = 1, \dots, m$), et $E[(Y_{ij} - e'_j \mu)^2 | \theta_j, j = 1, \dots, m]$ sont fixes et connues.

Remarque 2 : L'estimateur proposé par Cohen et Sommers (1984) est un cas particulier de $\hat{\theta}_{ij}$ lorsque

$$V_j = \left(\sum_{i=1}^I W_{ij} \right)^{-1} \text{Diag}(\sigma_1^2, \dots, \sigma_I^2) \cdot$$

Notons que ces auteurs supposent qu'il n'existe pas de corrélation entre les strates de produits et services, ce qui semble être une hypothèse très restrictive du point de vue de notre étude.

Remarque 3 : Il convient de souligner que par une simple inversion de matrice (voir Rao 1973),

$$V_j - \left(\sum_{m=1}^j V_j^{-1} \right)^{-1} = V_j \left[V_j + \left(\sum_{s=1}^s V_s^{-1} \right)^{-1} \right]^{-1} V_j$$

qui est définie positive. De plus,

$$E[(Y_{ij} - e'_j \mu)^2 | \theta_j, j = 1, \dots, m] = e'_j V_j \left[V_j + \left(\sum_{s=1}^s V_s^{-1} \right)^{-1} \right]^{-1} V_j e_j$$

$$+ \left[\theta_{ij} - e'_j \left(\sum_{m=1}^j V_j^{-1} \right)^{-1} \left(\sum_{m=1}^j V_j^{-1} \theta_j \right) \right]^2 \cdot$$

Par ailleurs, lorsque $\theta_j = \mu$, $\hat{a}_{ij} = 1$ et, par conséquent, $\hat{\theta}_{ij} = e'_j \mu$. Autrement, la grandeur du facteur de rétrécissement dépend de la valeur de

$$\left[\theta_{ij} - e'_j \left(\sum_{m=1}^j V_j^{-1} \right)^{-1} \left(\sum_{m=1}^j V_j^{-1} \theta_j \right) \right]^2 \cdot$$

Plus l'écart entre θ_{ij} et $e'_j (\sum_{m=1}^j V_j^{-1})^{-1} (\sum_{m=1}^j V_j^{-1} \theta_j)$ est grand, plus \hat{a}_{ij} est faible. Autrement dit, si une région particulière se distingue nettement des autres par ses caractéristiques, notre méthode attribuera moins de poids à la portion synthétique de l'estimateur. C'est ce qui explique la forte variabilité des facteurs de rétrécissement du tableau 1.

Tableau 1

Facteurs de rétrécissement \hat{a}_{ij} pour la région non autoreprésentative de l'Ouest

$i \setminus j$	1	2	2
1	0.8479225	0.7057626	0.9214804
2	0.8434894	0.5692695	0.8092725
3	0.0969009	0.0786758	0.6953904
4	0.4446537	0.5444809	1
5	0.6999551	0.3460123	0.5487382
6	0.0318442	0.4981756	0.2598752

nombre total d'unités de consommation dans la région de publication j . L'estimateur préliminaire de θ_{ij} est défini par l'expression

$$(2.1) \quad Y_{ij} = \sum_{n_j}^{l=1} w_{jl} X_{ijl} / \sum_{n_j}^{l=1} w_{jl}, \quad (i = 1, \dots, I; j = 1, \dots, m).$$

De même, l'estimateur correspondant pour une région principale s'écrit

$$(2.2) \quad Y_i = \sum_{n_j}^{j=1} \sum_{n_j}^{l=1} w_{jl} X_{ijl} / \sum_{n_j}^{j=1} \sum_{n_j}^{l=1} w_{jl}.$$

Comme la variabilité de X_i est beaucoup moins grande que celle de X_{ij} , il nous faut un estimateur composite de θ_{ij} qui accroisse le niveau de précision. Soit $X_j = (X_{1j}, \dots, X_{Ij})'$ et $\theta_j = (\theta_{1j}, \dots, \theta_{Ij})'$, $j = 1, \dots, m$. Posons V_j comme la matrice de variances-covariances vraie de X_j , ($j = 1, \dots, m$). Suivant une hypothèse artificielle, à savoir $\theta_j = \mu$, un vecteur colonne $I \times 1$, ($j = 1, \dots, m$), le meilleur estimateur de θ_j est défini

$$(2.3) \quad \bar{\mu} = \left(\sum_{m}^{j=1} V_j^{-1} \right) \left(\sum_{m}^{j=1} V_j^{-1} X_j \right),$$

expression que l'on obtient en minimisant $\sum_{m}^{j=1} (X_j - \mu)' V_j^{-1} (X_j - \mu)$ par rapport à μ . Or, l'hypothèse artificielle se vérifie difficilement. Par contre, lorsqu'il n'y a aucune espèce de similitude entre les θ_j , il convient d'utiliser X_j comme estimateur de θ_j . Pour les cas intermédiaires, on peut se servir d'un estimateur composite comme celui-ci:

$$(2.4) \quad \hat{\theta}_{ij}(a_{ij}) = (1 - a_{ij}) X_{ij} + a_{ij} e_j' \bar{\mu},$$

où les a_{ij} sont des valeurs constantes ($0 \leq a_{ij} \leq 1$), et e_j est un vecteur colonne $I \times 1$ dont l'élément i est égal à 1 et les autres éléments sont nuls.

On obtient a_{ij} en minimisant l'erreur quadratique moyenne

$$(2.5) \quad E \{ [(1 - a_{ij}) X_{ij} + a_{ij} e_j' \bar{\mu} - \theta_{ij}]^2 \mid \theta_{ij} \}$$

par rapport à a_{ij} . La formule optimale est donc

$$(2.6) \quad a_{ij} = \frac{E \{ (X_{ij} - e_j' \bar{\mu})^2 \mid \theta_{ij}, j = 1, \dots, m \}}{e_j' \left[V_j - \left(\sum_{m}^{j=1} V_j^{-1} \right)^{-1} e_j \right]}.$$

Par conséquent, l'estimateur optimal de θ_{ij} , pour la classe d'estimateurs définis en (2.4), s'écrit

$$(2.7) \quad \hat{\theta}_{ij} = (1 - a_{ij}) X_{ij} + a_{ij} e_j' \bar{\mu}.$$

relative. Les estimations finales du poids en valeur pour la catégorie de dépenses et pour la région principale sont égales aux estimations provisoires correspondantes. Si nous prenons les moyens d'obtenir par un procédé de balayage des données qui concordent, c'est probablement parce que, aux niveaux d'agrégation supérieurs les estimateurs préliminaires ont en général un rendement plus satisfaisant qu'aux niveaux d'agrégation inférieurs. La dernière étape de la méthode du BLS consiste à déduire directement des estimateurs finals du poids en valeur les estimateurs finals de l'importance relative pour chaque division.

Contrairement à d'autres auteurs, nous tenons compte de la corrélation entre les strates de produits et services lorsque nous présentons dans la section 2 nos estimateurs composites. Le facteur de rétrécissement de l'estimateur composite qui présente l'erreur quadratique moyenne la moins élevée d'une catégorie appropriée d'estimateurs comprend des paramètres inconnus. Ceux-ci sont estimés au moyen des échantillons répétés compensés de l'enquête. L'estimateur que proposent Cohen et Sommers (1984) est en fait un cas particulier de notre estimateur dans l'hypothèse où il n'existe aucune corrélation entre les estimateurs préliminaires.

Dans la section 2, nous cherchons essentiellement à estimer le poids en valeur d'une strate de produits et services pour une région de publication. Il est toutefois possible d'estimer des poids en valeur pour des niveaux d'agrégation plus élevés (par ex. : catégorie de dépenses pour une région de publication, etc.) en effectuant les sommations voulues. D'après les résultats de notre étude, si l'on en juge par l'erreur quadratique moyenne, les estimateurs que nous proposons surclassent constamment les estimateurs préliminaires correspondants et, donc, les estimateurs du BLS (notons qu'à cause du procédé de balayage, les estimateurs du BLS sont identiques aux estimateurs préliminaires pour les niveaux d'agrégation supérieurs).

Dans la section 3, nous proposons un estimateur composite de l'importance relative d'une strate de produits et services pour les régions de publication. Au lieu d'utiliser les estimateurs préliminaires des poids en valeur, nous servons des estimateurs préliminaires de l'importance relative de chacune des strates de produits et services qui constituent la catégorie de dépenses à l'étude. Comme la somme de ces estimateurs est égale à un, la matrice de variances-covariances des estimateurs préliminaires est singulière, ce qui fait que l'estimation de l'importance relative est un problème différent de l'estimation des poids en valeur. La méthode que nous proposons permet de supprimer de façon optimale une strate de produits et services, ce qui élimine le problème de la singularité de la matrice de variances-covariances des estimateurs préliminaires. Les résultats numériques de notre analyse montrent que l'estimateur proposé surclasse invariablement tous les autres estimateurs étudiés en ce qui regarde l'erreur quadratique moyenne. La section 4 renferme tous les résultats numériques. Nous avons évalué différents estimateurs du poids en valeur et de l'importance relative en nous fondant sur l'erreur quadratique moyenne estimée établie à l'aide des demi-échantillons compensés (voir McCarthy (1969), Ghosh et Sohn (1990)). D'après nos résultats, les estimateurs proposés semblent être supérieurs à tous les autres estimateurs étudiés.

2. ESTIMATION DU POIDS EN VALEUR

Soit X_{ijl} la moyenne des dépenses de deux semaines consécutives faites pour tous les produits ou services de la strate i par l'unité de consommation l dans la région de publication j d'une région principale particulière ($i = 1, \dots, I; j = 1, \dots, m; l = 1, \dots, n_j$). Soit W_{ijl} le poids d'échantillonnage rattaché à l'unité de consommation l dans la région de publication j ($j = 1, \dots, m; l = 1, \dots, n_j$). Ce poids représente un certain nombre d'unités de consommation dans la population et est déterminé par le Census Bureau selon une méthode complexe qui tient compte de divers facteurs comme les probabilités de sélection, la non-réponse, etc. Le but de cette section est d'estimer θ_{ij} , c.-à-d. les dépenses hebdomadaires moyennes réelles d'une unité de consommation pour les produits ou services de la strate i dans la région de publication j . Le poids en valeur est défini simplement comme $N_j \theta_{ij}$, N_j étant le

P^t_i = le prix moyen de tous les produits ou services de la strate i à la période t ($t = 0, T$); $\bar{Q}^{t_0}_i$ = la quantité totale de biens ou de services de la strate i achetés à la période $t = 0$ (période de référence).

Alors, l'indice de Laspeyres au temps $t = T$ est défini

$$I_T = \frac{\sum_{i \in E} \bar{Q}^{t_0}_i P^{iT}_i}{\sum_{i \in E} \bar{Q}^{t_0}_i P^{i0}_i} = \frac{\sum_{i \in E} C_i}{\sum_{i \in E} C_i (P^{iT}/P^{i0})} = \sum_{i \in E} R_i (P^{iT}/P^{i0}),$$

où

$$C_i = \bar{Q}^{t_0}_i P^{i0}_i = \text{dépenses totales pour les produits ou services de la strate } i \text{ à } t = 0, \\ R_i = C_i / \sum_{i \in E} C_i = \text{proportion de la dépense totale consacrée aux produits ou services de la strate } i \text{ à } t = 0.$$

Les quantités C_i et R_i représentent respectivement le "poids en valeur" et l' "importance relative" de la strate de produits et services i dans la catégorie de dépenses E .

Le Bureau of Labor Statistics se sert des données de la Consumer Expenditure Survey (CES) des E.-U. pour calculer les indices des prix à la consommation. La CES est constituée de deux volets: une enquête journal et une enquête par interview, chacune ayant ses plans d'échantillonnage et ses questionnaires propres. Dans cet article, nous ne considérons que des données de l'enquête journal. Selon le plan de sondage, toutes les unités primaires d'échantillonnage (UPÉ) d'une région autorenseignante particulière sont tirées avec une probabilité égale à un, mais dans le cas des régions non autorenseignantes, on ne tire qu'un échantillon probabiliste d'UPÉ. Dans un deuxième temps, on prélève un échantillon d'unités de consommation (UC) dans chacune des UPÉ échantillonnées en utilisant une fois de plus un plan probabiliste. Pendant deux périodes consécutives d'une semaine, chaque répondant consigne dans un journal les dépenses qu'il fait pour divers produits. Pour en savoir plus sur l'IPC et la CES, le lecteur est prié de consulter le *Handbook of Methods* du BLS (1988).

Les estimateurs classiques du poids en valeur et de l'importance relative des strates de produits et services sont généralement beaucoup moins efficaces au niveau de la région de publication qu'aux niveaux d'aggrégation géographique supérieurs (par ex., régions principales). Cela s'explique par le fait que les régions de publication ne comptent chacune qu'un petit nombre d'unités de consommation. D'où la nécessité d'améliorer les estimateurs classiques en "empruntant" de l'information à des sources connexes. Marks (1978) et Cohen et Sommers (1984) ont étudié certains estimateurs composites qui groupent de l'information en provenance de régions connexes. Ghosh et Sohn (1990) ont obtenu des estimateurs composites du poids en valeur et de l'importance relative à l'aide d'une méthode empirique de Bayes. La méthode qu'utilise actuellement le Bureau of Labor Statistics comporte plusieurs étapes. Premièrement, on calcule des estimateurs composites de l'importance relative à l'aide d'une méthode proposée par Cohen et Sommers (1984). On se sert ensuite d'une méthode itérative de "balayage" pour déduire les estimateurs du poids en valeur des estimateurs de l'importance

Une méthode multivariée pour l'estimation composite
des dépenses de consommation en vue du calcul
des indices des prix à la consommation
aux États-Unis

P. LAHIRI et WENYU WANG¹

RÉSUMÉ

Les auteurs se penchent sur l'estimation du "poids en valeur" et de l'"importance relative" de diverses strates de produits et services pour les régions de consommation. L'estimation de ces paramètres est une opération indispensable pour la construction des indices des prix à la consommation aux E.-U. Dans cet article, on se sert de modèles à plusieurs variables pour construire des estimateurs composites qui intègrent de l'information provenant de sources pertinentes. L'erreur quadratique moyenne (EQM) des estimateurs proposés et des estimateurs existants est estimée au moyen des demi-échantillons répétés tirés de l'enquête. D'après les résultats obtenus, les estimateurs proposés semblent être supérieurs aux autres estimateurs.

MOTS CLÉS: Dépenses de consommation; estimation composite; indice des prix à la consommation; poids en valeur; enquête journal; demi-échantillon; indice de Laspeyres; erreur quadratique moyenne; estimation synthétique.

1. INTRODUCTION

L'indice des prix à la consommation (IPC) des E.-U. est un indicateur des variations de prix touchant un ensemble de biens et services dont le volume et la qualité sont constants pendant une période quelconque. À chaque mois, le Bureau of Labor Statistics (BLS) des E.-U. calcule un certain nombre d'indices des prix à la consommation pour diverses régions géographiques, unités de consommation et classes de produits (voir le BLS Handbook of Methods, 1988). La plus petite unité de classification des produits pour laquelle le BLS calcule des IPC est la "strate de produits et services". Il s'agit d'un ensemble préalable de biens ou services de consommation (par ex.: lait entier frais) qui peuvent être achetés au détail durant une "période de référence" par un ensemble déterminé d'unités de consommation. Une unité de consommation peut être formée de tous les membres d'un ménage qui sont unis entre eux par les liens du sang, du mariage ou de l'adoption ou par toute autre forme de contrat. Un groupe de strates de produits et services forme ce que l'on appelle une catégorie de dépenses (par ex.: produits laitiers). À des fins d'échantillonnage, on a divisé les E.-U. en huit régions dites "principales". Une région principale peut être "autoreprésentative" ou "non autoreprésentative" et fait partie de l'une ou l'autre des quatre régions géographiques suivantes: Nord-Est, Midwest, Sud et Ouest. Une région autoreprésentative se compose de toutes les grandes villes d'une région. Une région non autoreprésentative correspond généralement à un comté ou à un groupe de comtés contigus. À des fins de publication, les régions principales sont subdivisées en "régions de consommation" ou "régions de publication".

Nous décrivons ci-dessous la formule de Laspeyres qu'utilise le BLS pour calculer l'IPC pour une région donnée et la catégorie de dépenses E par exemple. Soit

¹ P. Lahiri, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588 0323, USA. Wenyu Wang, SUNY Health Science Center at Brooklyn, Box 1203, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.

BIBLIOGRAPHIE

- BROWN, P. R., et BISHOP, G. F. (1982). Who refuses and resists in telephone surveys? Some new evidence. Document présenté au MAPOR Annual Conference.
- CANNELL, C. F. (1964). Factors affecting the refusal rate in interviewing. Ann Arbor: Survey Research Center (document de travail non publié).
- CIALDINI, R. B. (1984). *Influence: The New Psychology of Modern Persuasion*. New York: Quill.
- CIALDINI, R. B. (1990). Deriving psychological concepts relevant to survey participation from the literatures on compliance, helping, and persuasion. Document présenté au Workshop on Household Survey Nonresponse, Stockholm.
- DURBIN, J., et STUART, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *Journal of the Royal Statistical Society, Series A*, 114, 163-206.
- GOWER, A. R. (1979). Non-response in the Canadian Labour Force Survey. *Technique d'enquête*, 5, 29-58.
- GOYDER, J. (1987). *The Silent Minority: Nonrespondents on Sample Surveys*. Boulder, CO: Westview Press.
- GROVES, R. M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.
- GROVES, R. M., CIALDINI, R. B., et COUPER, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly* (à paraître).
- GROVES, R. M., et FULTZ, N. H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.
- GROVES, R. M., et KAHN, R. L. (1979). *Surveys by Telephone*. New York: Academic Press.
- HAWKINS, D. F. (1975). Estimation of nonresponse bias. *Sociological Methods and Research*, 3, 461-488.
- HERZOG, A. R., et RODGERS, W. L. (1988). Age and response rates to interview sample surveys. *Journals of Gerontology*, 43, S200-S205.
- HOUSE, J. S., et WOLF, S. (1978). Effects of urban residence on interpersonal trust and helping behavior. *Journal of Personality and Social Psychology*, 36, 1029-1043.
- INDERFURTH, G. P. (1972). Investigation of Census Bureau interviewer characteristics, performance and attitudes: A summary. U.S. Bureau of the Census: Working Paper 34.
- KORTE, C., et KERR, N. (1975). Responses to altruistic opportunities in urban and nonurban settings. *Journal of Social Psychology*, 95, 183-184.
- LIEVESLEY, D. (1988). Unit non-response in interview surveys. London: Social and Community Planning Research (document de travail non publié).
- PAUL, E. C., et LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Techniques d'enquête*, 8, 48-85.
- RAUTA, I. (1985). A comparison of the census characteristics of respondents and nonrespondents to the 1981 General Household Survey (GHS). *Statistical News*, 71, 12-15.
- SCHYBERGER, B. W. (1967). A study of interviewer behavior. *Journal of Marketing Research*, 4, 32-35.
- SINGER, E., FRANKEL, M. R., et GLASSMAN, M. B. (1983). The effect of interviewer characteristics and expectation on response. *Public Opinion Quarterly*, 47, 68-83.
- SMITH, T. W. (1983). The hidden 25 percent: An analysis of nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly*, 47, 386-404.
- STEEH, C. G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.
- U.S. BUREAU OF THE CENSUS (1988). *County and City Data Book*, 1988. U.S. Government Printing Office.
- WEAVER, C. N., HOLMES, S. L., et GLENN, N. D. (1975). Some characteristics of inaccessible respondents in a telephone survey. *Journal of Applied Psychology*, 60, 260-262.
- WILCOX, J. B. (1977). The interaction of refusal and not-at-home sources of nonresponse bias. *Journal of Marketing Research*, 14, 592-597.

Efficacité: On a demandé aux intervieweurs dans quelle mesure ils étaient ou non d'accord avec l'énoncé suivant: Avec suffisamment d'efforts, je peux convaincre même le répondant le moins enthousiaste de participer à l'enquête. Echelle ordinale à quatre valeurs, 1 = fortement d'accord, 4 = fortement en désaccord. Une note élevée indique une plus grande confiance en soi.

Comportements de l'intervieweur

Autorité: On a demandé aux intervieweurs à quelle fréquence ils laissaient divers documents (demande de rendez-vous, exemplaire de la lettre de présentation, etc.) au domicile des répondants quand ils ne trouvaient personne à la maison. Les réponses à ces questions ont été combinées pour former une échelle de fréquence d'utilisation de documents servant à accroître l'autorité. Une note élevée montre une plus grande utilisation de l'autorité.

Echange: Cherchez-vous à compléter un élément du domicile ou de l'apparence personnelle du répondant?

1 = Toujours, parfois
0 = Rarement, jamais

Validation sociale: Dites-vous "La majorité des personnes aiment répondre à l'interview?"

1 = Toujours, parfois
0 = Rarement, jamais

Pertinence: Expliquez-vous aux répondants comment les résultats de l'enquête pourraient les toucher personnellement?

1 = Toujours, parfois
0 = Rarement, jamais

Rareté: Dites-vous à un répondant que l'interview doit être terminée pour une certaine date?

1 = Toujours, parfois
0 = Rarement, jamais

Cohérence: Avant qu'un répondant ait montré un signe quelconque de collaboration, commencez-vous à poser les questions de l'enquête?

Répertoire: Dans une question ouverte, on a demandé aux intervieweurs de fournir la liste de tout ce qu'ils font habituellement pour persuader un répondant peu enthousiaste de participer à l'enquête. Le nombre d'éléments uniques mentionnés sert d'indicateur du répertoire de techniques disponibles.

Adaptation: Pour une série de 15 éléments relatifs au comportement, les intervieweurs indiquaient s'ils avaient toujours, parfois, rarement ou jamais un tel comportement. On obtient un indicateur de l'adaptation à l'application de diverses techniques de persuasion en comptant le nombre de fois où un intervieweur a précisé une des catégories du milieu (parfois ou rarement) en réponse à ces questions. Une note élevée montre une plus grande utilisation de l'adaptation.

Zone d'affectation	
Densité de la population:	Densité de la population (personnes au mille carré).
Taux de criminalité:	Taux de criminalité (crimes pour 100,000 habitants).
Pourcentage de personnes de 65 ans ou plus:	Pourcentage de la population âgée de 65 ans ou plus.
Pourcentage de personnes de moins de 5 ans:	Pourcentage de la population âgée de moins de 5 ans.
Taille du ménage:	Taille moyenne du ménage.
Enquête	
Ensemble de variables fictives servant à désigner l'enquête à laquelle chaque intervieweur est affecté:	
HIS: L'intervieweur est-il affecté à la Health Interview Survey?	1 = Oui 0 = Non
NCS: L'intervieweur est-il affecté à la National Crime Survey?	1 = Oui 0 = Non
CE: (Consumer Expenditure Survey), cette enquête constitue donc la catégorie omise.	
Expérience de l'intervieur	
Durée des fonctions:	Mesurée en jours de travail à l'emploi du Census Bureau à titre d'intervieur, valeur transformée en années fractionnaires.
Etendue de l'expérience: Total du nombre d'organismes d'enquête différents pour lesquels l'intervieur a travaillé.	
Attentes de l'intervieur	
Confidentialité: On a demandé aux intervieweurs s'ils pensaient qu'il y avait des situations dans lesquelles le Census Bureau donnerait des renseignements fournis par un répondant dans le cadre d'une enquête à un ou plusieurs des organismes suivants (FBI, CIA, INS, IRS, organismes des administrations publiques d'Etat et locales).	
1 = Respect d'une confidentialité élevée (le Census Bureau ne donnerait de renseignements à aucun de ces organismes).	0 = Non-respect d'une confidentialité élevée (le Census Bureau donnerait des renseignements à un ou plusieurs de ces organismes).
Taux/qualité: Compromis entre le taux de réponse et la qualité des données. Lequel des énoncés ci-après décrit le mieux votre sentiment à titre d'intervieur:	
1 = Il est préférable de persuader un répondant peu enthousiaste de participer que d'accepter un refus.	0 = Il est préférable d'accepter un refus de la part d'un répondant peu enthousiaste.

6. DISCUSSION

Dans cet article on a cherché à mesurer si les intervieweurs expérimentés obtiennent des taux de réponse plus élevés que les intervieweurs inexpérimentés. On a trouvé que c'est le cas. On a ensuite essayé d'expliquer pourquoi cela se produit. Dans une large mesure, cette tentative a échoué. Une des raisons qui explique cette situation peut être le fait que le modèle est inexact. Cependant, les discussions qui se sont poursuivies avec les intervieweurs et le personnel de surveillance nous amènent à croire que cette formulation théorique a une certaine valeur.

On peut avancer quatre explications. Premièrement, on fait l'essai du modèle à un niveau d'aggrégation qui n'est pas approprié. Bien que le questionnaire se concentrait sur ce que les intervieweurs font généralement, nous sommes plus intéressés à connaître la façon dont ils agissent dans des situations particulières. Un test plus approprié portant sur ces idées devrait être réalisé au niveau du répondant ou du ménage. Deuxièmement, la mesure de divers concepts peut être inadéquate. On peut améliorer l'adaptation des concepts utilisés dans les documents sur l'acquiescement en comportements particuliers des intervieweurs. Troisièmement, il faut faire remarquer, à nouveau, que ces modèles portent sur les taux de réponse et non sur les taux de refus. Il se peut que certains comportements soient plus appropriés pour persuader des personnes dans l'échantillon de participer à l'enquête (on vise alors à réduire le nombre de refus), alors qu'il est possible que d'autres servent plus à réussir à accéder à des personnes dans l'échantillon (la partie de la non-réponse relative à l'impossibilité d'entrer en communication avec le répondant). On n'a pu élaborer ici des modèles distincts pour ces deux processus. Finalement, il est possible que d'autres caractéristiques non mesurées des intervieweurs (apparence, qualité de la voix, tenue, etc.) jouent aussi un rôle qui influence la décision du répondant. Ces lacunes possibles n'annulent pas le rôle de ces comportements pour ce qui est d'influencer les taux de réponse. Les conclusions suggèrent plutôt d'autres recherches et analyses pour examiner les rapports entre des comportements particuliers et leur application d'une part et les taux de réponse au niveau de l'intervieweur d'autre part. Nous pensons que ce champ d'enquête a une certaine valeur et travaillons en vue d'acquérir une meilleure compréhension du rôle de l'expérience, des attentes et du comportement des intervieweurs relativement à la participation à une enquête.

REMERCIEMENTS

Ce travail a été appuyé par le Bureau of the Census, le Bureau of Labor Statistics, le Bureau of Justice Statistics et le National Center for Health Statistics. Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles du Bureau of the Census ou de tout autre organisme. Les auteurs désirent remercier Lorraine McCall de son aide pour cette recherche. Nous remercions aussi les critiques pour leurs suggestions utiles.

ANNEXE A

VARIABLES UTILISÉES DANS LES ANALYSES

Voici un résumé des variables utilisées dans les analyses. On peut obtenir des exemplaires du questionnaire en s'adressant aux auteurs.

Variable dépendante

Taux de réponse: Il s'agit du taux de réponse obtenu par chaque intervieweur au cours de la période de six mois sur laquelle portait l'étude, exprimé en pourcentage.

Nous avons aussi fait des tests portant sur les interactions entre les trois enquêtes et diverses caractéristiques des tâches. Aucune de ces caractéristiques ne semble avoir un effet perceptible dans ces modèles et nous n'en traiterons pas davantage ici. Comme autre test pour la présence d'interactions additionnelles auxquelles participent les variables d'enquête, nous avons ajusté des modèles distincts pour chacune des trois enquêtes. Les modèles obtenus sont essentiellement les mêmes pour chacune des trois enquêtes étudiées. Ainsi, bien que le niveau de réponse diffère parmi les trois enquêtes, l'incidence **relative** de la durée des fonctions sur les taux de réponse semble être identique.

Compte tenu du fait qu'il semble que les intervieweurs expérimentés obtiennent des taux de réponse plus élevés peu importe la zone à laquelle ils sont affectés, nous pouvons maintenant nous attaquer à la question de savoir **comment** l'expérience a une incidence sur les niveaux de collaboration. Qu'est-ce qui rend un intervieweur plus expérimenté plus en mesure d'obtenir la collaboration des répondants?

La première étape comprend l'ajout, au modèle 2, de variables sur les attentes des intervieweurs. Les résultats sont présentés sous le titre Modèle 3 au tableau 2. Les trois variables des attentes vont toutes dans la direction prévue, bien qu'une seule soit statistiquement significative aux niveaux traditionnels. Il semble que les intervieweurs qui croient plus à leur aptitude à convaincre les répondants peu disposés à répondre de participer à l'enquête obtiennent effectivement des taux de réponse plus élevés.

Il faut signaler que le lien de causalité entre les attentes et les taux de réponse ne peut être établi dans une étude transversale comme celle-ci. Il se peut qu'un plus grand succès amène de plus grandes attentes d'un succès à venir, plutôt que le contraire. Cette interprétation va à l'encontre de l'hypothèse selon laquelle le fait d'insuffler une plus grande confiance en soi parmi les intervieweurs permettrait d'obtenir des niveaux de réponse plus élevés. Néanmoins, il s'agit d'une conclusion fascinante que l'on doit étudier plus à fond.

L'étape suivante consistait à ajouter au modèle l'ensemble de comportements des intervieweurs. On peut voir les résultats dans le modèle 4 du tableau 2. Il faut remarquer deux points à propos de ces résultats. Tout d'abord, l'explication fournie par l'inclusion de cet ensemble de comportements des intervieweurs n'a pas permis d'éliminer l'effet de la durée des fonctions. En fait, l'ajout soit des variables sur les attentes, soit des variables sur le comportement n'a guère d'effet sur le coefficient correspondant à la durée des fonctions.

Deuxièmement, les résultats pour les comportements particuliers sont un peu contradictoires. On s'attendait à ce que les coefficients pour toutes les variables de comportement soient positifs. Cela n'est pas le cas. Les résultats pour l'autorité et l'échange montrent que les intervieweurs qui emploient ces techniques obtiennent des taux de réponse plus élevés. Par contre, l'utilisation du principe de rareté semble avoir l'effet contraire. Il se peut bien que le fait d'exercer des pressions sur un répondant pour respecter certaines échéances se retourne contre l'intervieweur. Les autres variables de comportement ne semblent pas avoir un effet significatif sur les taux de réponse obtenus par les intervieweurs du Census Bureau.

On a proposé plus tôt de considérer l'utilisation d'un modèle réduit dans lequel on n'emploie que les variables "répétition" et "adaptation". Dans le tableau 2, on a vu que ces deux variables n'ont pas d'effets significatifs en présence des autres variables de comportement. Même après avoir supprimé du modèle les autres variables de comportement, les variables "répétition" et "adaptation" ont encore peu d'incidence sur les taux de réponse. Ainsi, ces données fournissent peu d'appui empirique à l'argument selon lequel la façon dont les intervieweurs utilisent diverses techniques d'acquisition est plus importante que les comportements réels eux-mêmes. Toutefois, il se peut que les mesures de ces deux concepts soient faibles et un meilleur test de leur rôle devrait être effectué au niveau du répondant.

Tableau 2
Résultats des analyses de régression par MCP des enquêtes NCS, HIS, CE taux de réponse au niveau des intervieweurs

	Modèle 1		Modèle 2		Modèle 3		Modèle 4	
	Coefficient	Err.-type	Coefficient	Err.-type	Coefficient	Err.-type	Coefficient	Err.-type
Ordonnée à l'origine	96.94	(3.19)	96.21	(5.39)	94.95	(3.25)	93.44	(3.35)
Zone d'affectation:								
Densité de la population	-0.00017**	(0.000023)	-0.000078*	(0.000038)	-0.000084*	(0.000038)	-0.000071	(0.000038)
Taux de criminalité	-0.00024**	(0.000055)	-0.00021**	(0.000055)	-0.00023**	(0.000056)	-0.00022**	(0.000056)
Pourcentage de personnes de 65 ans ou plus	-0.057	(0.051)	-0.054	(0.050)	-0.061	(0.051)	-0.061	(0.052)
Pourcentage de personnes de moins de 5 ans	0.41*	(0.16)	0.37*	(0.16)	0.29	(0.17)	0.35*	(0.17)
Taille du ménage	-3.20*	(1.70)	-2.92*	(1.24)	-2.88*	(1.26)	-3.09*	(1.27)
Indicateurs d'enquête:								
NCSI	6.72**	(0.40)	6.67**	(0.40)	6.68**	(0.41)	6.55**	(0.42)
HISI	5.65**	(0.46)	5.63**	(0.46)	5.64**	(0.47)	5.65**	(0.48)
Expérience des intervieweurs:								
Log (durée des fonctions)	0.62**	(0.14)	0.74**	(0.14)	0.69**	(0.15)	0.72**	(0.15)
Log (durée des fonctions)* densité			-0.00010**	(0.000032)	-0.00011**	(0.000032)	-0.00011**	(0.000032)
Attentes des intervieweurs:								
Confidentialité					0.61	(0.37)	0.59	(0.37)
Taux/qualité					0.046	(0.40)	-0.00073	(0.41)
Efficacité					0.55**	(0.15)	0.53**	(0.15)
Comportements des intervieweurs:								
Autorité							0.14**	(0.055)
Echange							0.67*	(0.29)
Validation sociale							0.18	(0.32)
Pertinence							-0.19	(0.33)
Rareté							-0.66*	(0.29)
Cohérence							-0.21	(0.29)
Répertoire							-0.0068	(0.065)
Adaptation							-0.042	(0.054)
R ² ajusté	0.3553		0.3640		0.3784		0.3873	
(n)	(679)		(679)		(645)		(639)	

** $p < .01$
* $p < .05$
1 Les intervieweurs travaillant à l'enquête CE constituent la catégorie omise.

5. RÉSULTATS

Nous avons tout d'abord mesuré l'incidence de l'expérience, en "neutralisant" les caractéristiques des zones d'affectation et les variables fictives utilisées pour les enquêtes (modèle 1 du tableau 2). Examinons tout d'abord les coefficients des variables de contrôle. Avec de rares exceptions, la majorité des variables relatives à la zone d'affectation ont une incidence significative sur les taux de réponse. Tant la densité de la population que le taux de criminalité ont l'effet prévu, on obtient des taux de réponse inférieurs dans les régions à criminalité élevée et à forte densité de population. L'effet négatif de la taille du ménage est contraire à ce que l'on attendait. Cette situation peut être expliquée, en partie, par le fait que ces enquêtes recueillent toutes des renseignements auprès ou à propos de **tous** les membres adultes du ménage, augmentant de ce fait le fardeau de déclaration pour les gros ménages. Cela est contraire à ce qui se produit dans de nombreuses enquêtes où un seul adulte est choisi dans chaque ménage. L'effet de l'âge correspond à notre hypothèse, les taux de réponse tendant à être inférieurs (mais pas de façon significative) dans les zones où il y a de plus grandes proportions de personnes de plus de 65 ans, mais plus élevés dans les régions où l'on retrouve de nombreux ménages avec de jeunes enfants.

Les effets considérables pour les deux variables d'enquête (par rapport à la catégorie omise de la Consumer Expenditure Survey) reflètent des différences dans les taux de réponse moyens pour ces trois enquêtes. De tels écarts peuvent être attribués à une foule de différences relatives au plan d'enquête (durée de l'interview, règles pour la sélection des répondants, plans par panel par opposition à plans transversaux, contenu des questionnaires, etc.) qui dépassent la portée du présent article. Néanmoins, il est clairement nécessaire de "neutraliser" ces différences.

Examinons maintenant l'effet mesuré de l'expérience, compte tenu de ces variables de contrôle. On peut voir que la durée des fonctions a un effet positif important sur les taux de réponse, même quand on "neutralise" la nature de la zone à laquelle un intervenieur est affecté. Cela semble confirmer les opinions courantes à propos du rôle de l'expérience des intervenieurs. Les différences entre les intervenieurs pour ce qui est des taux de réponse semblent être plus que de simples artefacts des différences dans les zones auxquelles ils sont affectés et l'expérience joue un rôle important dans de telles différences entre les intervenieurs.

Nous avons aussi fait des tests portant sur l'inclusion d'un indicateur pour l'étendue de l'expérience, mais nous avons trouvé qu'un tel indicateur n'a pas d'effet significatif en présence des variables restantes. Donc, pour les intervenieurs du Census Bureau du moins, il semble que l'expérience acquise lors du travail effectué pour d'autres organismes réalisant des enquêtes n'ait pas, sur les taux de réponse, d'incidence marginale qui dépasse celle de la durée des fonctions.

La durée des fonctions a-t-elle une incidence différentielle sur les taux de réponse dans différentes zones d'affectation? Le modèle 2 du tableau 2 inclut un terme d'interaction entre le logarithme de la durée des fonctions et la densité de la population. On a aussi fait un test portant sur un terme additionnel d'interaction entre la durée des fonctions et le taux de criminalité, mais on a trouvé que ce coefficient n'était pas significatif et que l'interaction avait peu d'incidence sur les autres éléments du modèle. Le terme d'interaction dans le modèle 2 est statistiquement significatif, mais son signe est le contraire de celui que l'on prévoyait. Nous avons fait l'hypothèse que l'expérience aurait une incidence plus élevée dans les régions à forte densité, mais il ne semble pas que ce soit le cas. Une autre explication pourrait être un "effet d'épuisement professionnel". Il se peut que les intervenieurs plus expérimentés travaillant dans les régions urbaines à forte densité de population perdent leur enthousiasme plus rapidement que les intervenieurs expérimentés employés dans les régions où le travail est moins difficile et cela contribue à des taux de réponse moins élevés. L'épuisement professionnel des intervenieurs peut être un facteur qui contribue aux taux de roulement plus élevés dans les grandes régions métropolitaines.

Tableau 1

Taux de non-réponse en 1990 pour trois enquêtes

Enquête	Taux de non-réponse	Taux de refus
Consomer Expenditure Survey (CE)	13.4	11.6
Health Interview Survey (HIS)	4.5	2.8
National Crime Survey (NCS)	3.1	1.6

On fait l'hypothèse que l'effet de la durée des fonctions sur le taux de réponse est plus élevé au cours des premières années d'emploi. La variable "durée des fonctions" est transformée (on utilise le logarithme naturel) afin de refléter cette hypothèse. La variable transformée produit effectivement une amélioration dans l'ajustement par rapport à la forme linéaire de la variable "durée des fonctions".

On peut trouver, à l'annexe A, une description plus détaillée des variables utilisées dans ces analyses.

4. RESTRICTIONS

Avant de décrire les analyses, il est important de faire remarquer certaines des restrictions s'appliquant à ces données. Tout d'abord, ces conclusions ne se rapportent qu'aux intervieweurs travaillant à trois enquêtes permanentes nationales du Census Bureau au moment où l'interview auprès des intervieweurs a été réalisée. On ne peut généraliser les conclusions à d'autres enquêtes téléphoniques ou directes réalisées par des organismes universitaires ou du secteur privé.

De plus, les données sont de nature transversale. Les effets de cohorte et de période sont confondus avec les effets de l'expérience. C'est-à-dire que toutes les différences observées dans les taux de réponse selon l'expérience de l'intervieweur peuvent être dues à des changements dans la qualité des intervieweurs embauchés au fil des ans, dans l'efficacité de la formation des intervieweurs dans le temps ou dans la rotation différentielle selon la qualité des intervieweurs. On peut construire des hypothèses afin d'appuyer les effets tant positifs que négatifs de ces facteurs sur les taux de réponse. L'incidence mesurée de l'expérience des intervieweurs sur les taux de réponse est donc une combinaison complexe de ces facteurs. On doit disposer de mesures longitudinales sur les intervieweurs pour démêler ces effets.

Les intervieweurs ne sont pas affectés aléatoirement à des zones. Bien que nous ayons tenté de "neutraliser" un certain nombre de caractéristiques des zones d'affectation qui peuvent avoir une incidence sur les taux de réponse, il peut exister de nombreux autres facteurs qui pourraient expliquer les différences dans les taux de réponse entre les zones d'affectation. De plus, nous sommes limités à des facteurs de contrôle que nous laissons à désirer sur les attributs des comités et des groupes de comités, pas sur les attributs de zones d'affectation particulières des comités. Une analyse hiérarchique renfermant des données sur des répondants particuliers et sur les personnes qui les ont interviewés améliorerait ces facteurs de contrôle.

Finalement, on a mesuré la variable dépendante pour une période qui va jusqu'à celle où l'on a fait remplir le questionnaire par les intervieweurs, inclusivement. On ne disposait pas de données plus récentes sur le taux de réponse à ce moment. Compte tenu du fait que l'on n'a pas mesuré les comportements et les attentes avant d'obtenir les taux de réponse, on doit faire preuve de prudence avant d'établir des rapports de cause à effet.

En dépit de ces restrictions, ces données nous fournissent l'occasion de vérifier les opinions courantes à propos du rôle de l'expérience des intervieweurs en ce qui a trait au taux de réponse et d'étudier le rôle des attentes et du comportement des intervieweurs lors d'enquêtes directes.

Il faut remarquer que ces variables ne peuvent refléter que des différences grossières dans la zone d'affectation d'un intervieweur et ne peuvent, par exemple, faire la distinction entre les régions du centre d'une ville et celles des banlieues.

La date à laquelle chaque intervieweur a été embauché par le Census Bureau a été obtenue à partir des dossiers administratifs afin de créer une variable utilisée comme mesure de la durée des fonctions. Bien que cette variable n'indique pas la durée de l'expérience pour une enquête particulière, elle reflète la période pendant laquelle un intervieweur a été employé par le Census Bureau.

Un inconvénient majeur de cette étude est qu'il n'a pas été possible d'obtenir des mesures de la race, de l'âge, du sexe ou d'autres attributs démographiques de l'intervieweur. Les restrictions en matière de confidentialité ne nous ont pas permis d'accéder aux dossiers personnels pour recueillir ces renseignements, pas plus que nous n'avons pu inclure ces sujets dans le questionnaire soumis aux intervieweurs.

3.3 Plan analytique

Trois enquêtes différentes sont représentées dans l'ensemble de données. Plutôt que d'introduire des variables de contrôle mesurant des caractéristiques fondamentales des plans employés pour les enquêtes, nous avons utilisé des variables auxiliaires pour "neutraliser" les différences importantes, au niveau conception, entre ces enquêtes.

La variable dépendante est le taux de réponse agrégé pour la période de six mois allant d'octobre 1989 à mars 1990. Il n'a pas été possible d'obtenir des données au niveau de l'intervieweur sur les composantes de la non-réponse (particulièrement sur les refus) pour cette période. Ces taux ne permettent donc pas de faire la distinction entre les composantes "impossibilité d'entrer en communication" et "refus" de la non-réponse. Il faut donc remarquer que les analyses présentées ici sont basées sur des taux de **réponse** au niveau de l'intervieweur plutôt que sur des taux de **refus**.

Les taux de non-réponse pour les trois enquêtes en 1990 (basés sur les totaux de l'échantillon national) sont donnés au tableau 1.

La proportion des refus par rapport à la non-réponse totale varie de 87% pour l'enquête CE à 52% pour l'enquête NCS. Nous soupçonnons que différents ensembles de facteurs agissent sur ces deux composantes de la non-réponse. Idéalement, des modèles distincts seraient ajustés pour chaque composante, mais cela n'a pas été possible, compte tenu des données courantes. Dans la mesure où les facteurs qui ont un effet sur les refus diffèrent de ceux qui ont un effet sur d'autres composantes de la non-réponse (comme l'impossibilité d'entrer en communication), il y aura confusion dans l'interprétation des résultats (voir Lievesley 1988). On peut aussi voir que les taux de non-réponse pour ces trois enquêtes sont faibles pour commencer. Cela peut limiter davantage l'aptitude de ces modèles à expliquer les différences entre les intervieweurs.

Compte tenu du fait que la taille des tâches des intervieweurs varie (et qu'elle a, par conséquent, un effet sur la variance de chaque taux de réponse mesuré), nous avons utilisé les moindres carrés pondérés (MCP) avec la taille de la tâche comme poids. Nous avons comparé les résultats des MCP à ceux obtenus à l'aide des moindres carrés ordinaires (MCO) et avons constaté que les MCP réduisent très légèrement la valeur des coefficients, mais qu'ils n'ont aucun effet sur le signe ou sur la valeur relative des coefficients. Toutes les analyses présentées ici sont basées sur les solutions obtenues à l'aide des MCP.

Nous avons effectué une série de tests afin de déterminer dans quelle mesure les modèles précisés sont appropriés. Nous avons détecté un certain nombre de valeurs aberrantes dans la variable dépendante. Toutefois, la suppression de ces valeurs aberrantes a eu peu ou pas d'effet sur les résultats obtenus et, par conséquent, nous avons conservé ces valeurs dans toutes les analyses. Nous avons aussi effectué des tests de l'hypothèse de normalité. Les courbes de probabilité normale montrent que les résidus provenant de ces modèles ne diffèrent pas beaucoup de ceux obtenus pour une distribution normale.

3. MÉTHODE

3.1 Stratégies de collecte des données

Les résultats présentés dans cet article font partie d'une étude plus considérable sur la participation à une enquête lors d'enquêtes directes aux États-Unis. Dans la première partie du travail on a eu recours à une série de groupes de décision dont faisaient partie des intervieweurs travaillant à une variété d'enquêtes différentes dans tout le pays. Les renseignements obtenus à l'aide de ces groupes ont mené à l'élaboration d'un questionnaire structuré visant à tester certaines de ces hypothèses auprès d'un nombre plus considérable d'intervieweurs. Les enquêtes portant sur les intervieweurs avaient pour objectif de mesurer les influences comportementales, expérimentelles et attitudinales sur les niveaux de collaboration obtenus par les intervieweurs. L'élaboration et l'essai du questionnaire ont été effectués par des employés du Survey Research Center en collaboration avec des employés du U.S. Census Bureau.

On a fait remplir ce questionnaire par des intervieweurs du U.S. Census Bureau travaillant aux trois enquêtes par interview sur place suivantes: a) la Consumer Expenditure Quarterly Survey (CE), commandée par le Bureau of Labor Statistics; b) la National Health Interview Survey (HIS), commandée par le National Center for Health Statistics et c) la National Crime Survey (NCS), commandée par le Bureau of Justice Statistics.

Le questionnaire a été posté, en février 1990, aux intervieweurs du Census Bureau travaillant à ces trois enquêtes. Tous les intervieweurs ont été payés à leur taux de rémunération habituel pour remplir le questionnaire (la majorité des intervieweurs ont reçu une rémunération correspondant à une heure de travail). Dans une tentative en vue d'obtenir des réponses sincères et pour éliminer la menace d'intervention de la part des surveillants, on a assuré les intervieweurs que leurs réponses particulières ne seraient pas vues par un quelconque de leurs surveillants, qu'on n'en parlerait pas avec ces derniers et que les résultats ne seraient publiés que sous forme de totaux statistiques.

Les questionnaires ont été retournés au bureau central par la poste. On a eu recours à des lettres et à des appels téléphoniques de rappel pour accroître le taux de réponse. On a reçu 1,013 questionnaires remplis, ce qui représente un taux de réponse de 97,1%. Un certain nombre des questionnaires ont été exclus des analyses présentées ici. Tous les intervieweurs surveillants (256) ont été exclus. Il arrive souvent que ces personnes n'ont pas, en propre, de tâches régulières et, généralement, elles travaillent à un certain nombre d'enquêtes différentes. On fait souvent appel à ces intervieweurs pour obtenir des réponses quand il y a eu refus ou pour terminer des tâches incomplètes. Si l'on exclut les intervieweurs surveillants, il est rare qu'il y ait transfert de tâche d'un intervieweur à un autre dans les enquêtes étudiées. Aux fins du calcul des taux de réponse au niveau de l'intervieweur, chaque cas de non-réponse a été attribué à l'intervieweur original, même si, par la suite, un autre intervieweur a pu obtenir la collaboration du répondant. De plus, les intervieweurs qui ont commencé à travailler pendant la période au cours de laquelle l'interview auprès des intervieweurs a été réalisée, et pour lesquels on ne disposait pas de données chronologiques sur leurs taux de réponse ont aussi été exclus (46 intervieweurs). Cela a laissé un total de 711 intervieweurs, 207 travaillant à l'enquête CE, 139 à l'enquête HIS et 365 à l'enquête NCS. Il se peut que le nombre de cas inclus dans les analyses soit réduit davantage à cause de données manquantes pour certaines variables.

3.2 Structure des données

En plus des réponses au questionnaire, d'autres variables ont été ajoutées au fichier de données. Elles comprenaient un ensemble de variables permettant de représenter la zone d'affectation de chaque intervieweur. Généralement, l'unité primaire d'échantillonnage (UPÉ) dans laquelle un intervieweur travaille est composée d'un comté ou de plusieurs comtés limitrophes. Les données au niveau du comté ont été tirées du *County and City Data Book* (Bureau of the Census 1988), groupées au niveau de l'UPÉ et jointes aux enregistrements des intervieweurs.

Premièrement, le problème que pose l'obtention de la collaboration des membres d'un échantillon dans les centre-villes est bien connu (voir Steeh 1981; Smith 1983). House et Wolf (1978) ont trouvé que les taux de criminalité croissants, particulièrement dans les régions urbaines à forte densité de population, ont fortement contribué à dissuader les personnes de participer à des enquêtes et ont nui au comportement de confiance et d'assistance en général (Korte et Kerr 1975). Nous supposons que cela découle à la fois de la répugnance des résidents à s'associer à des étrangers et de l'inquiétude qu'ont les intervieweurs à pénétrer dans ces quartiers.

Si l'on se reporte aux caractéristiques des ménages faisant partie d'un échantillon, on a trouvé qu'il existait une corrélation positive entre la taille des ménages et les taux de réponse (voir Gower 1979; Paul et Lawes 1982; Rautra 1985). Pour les ménages composés d'une seule personne, les taux de refus tendent à être relativement élevés (voir Brown et Bishop 1982; Wilcox 1977). Cela peut être dû, en partie, à la proportion élevée de personnes âgées qui vivent seules. Par contre, pour les familles avec des enfants à charge, les taux de réponse tendent à être plus élevés. Selon Lievesley (1988), la probabilité élevée de trouver quelqu'un à la maison découle de proportions élevées d'enfants âgés de 0 à 4 ans peut expliquer des taux de réponse plus élevés dans certaines régions du R.-U.

Les résultats relatifs aux caractéristiques des personnes faisant partie d'un échantillon sont un peu moins clairs. Un certain nombre de chercheurs (voir Brown et Bishop 1982; Hawkins 1975; Herzog et Rogers 1988; Weaver 1975) ont trouvé qu'il y avait une association entre l'âge et la non-réponse. L'incidence d'autres caractéristiques des personnes dans un échantillon comme la race, la scolarité, le statut socio-économique, le sexe, *etc.* est quelque peu contradictoire (voir Groves (1989) et Goyder (1987) pour des études de ces facteurs).

2.5 Caractéristiques des plans d'enquête

Finalement, les caractéristiques des plans d'enquête (sujet, fardeau de déclaration, règles pour le choix des répondants, *etc.*) ont vraisemblablement une influence sur la décision, que prend une personne faisant partie d'un échantillon, de participer à l'enquête correspondante, à la fois directement et à cause des contraintes relatives aux attentes et au comportement des intervieweurs.

2.6 Effets des interactions sur le taux de réponse

Nous soupçonnons qu'il peut y avoir un certain nombre d'effets d'interactions statistiques qui ont des influences sur la non-réponse. Une question est de savoir s'il existe certaines régions (comme les régions des centre-villes à forte densité de population) dans lesquelles l'expérience d'un intervieweur est plus importante que dans d'autres. Par exemple, il se peut que les régions urbaines à forte densité de population puissent être plus diverses, exigeant une plus grande expérience pour faire face à une plus grande variété de situations différentes. Le comportement dans des régions où les situations auxquelles les intervieweurs ont à faire face sont toutes fort semblables pourrait être plus facile à apprendre, puisqu'on aurait besoin de moins de stratégies de persuasion.

Nous avons aussi le sentiment qu'il se peut que, pour des enquêtes différentes, on obtienne des taux de réponse qui varient pour des sous-populations différentes parce que l'importance du sujet de l'enquête n'est pas la même pour de tels groupes. Par exemple, on peut s'attendre à ce que, pour la National Crime Survey (qui se concentre sur la victimisation criminelle), on obtienne de meilleurs taux de réponse dans les régions où la criminalité est élevée que dans celles où la criminalité est faible. De même, pour la National Health Interview Survey (qui mesure des activités reliées à la santé) on peut obtenir des taux de réponse plus élevés dans les régions où la population est plus âgée que la moyenne. On peut s'attendre à des interactions semblables entre la Consumer Expenditure Survey et les variables telles que la taille moyenne des ménages et le niveau moyen de revenu.

d) Autorité: On accepterait mieux de céder aux demandes d'une personne que l'on considère être une autorité légitime.

e) Rareté: On accepterait mieux d'acquiescer à des demandes afin de profiter d'occasions qui sont rares.

f) Estime: On accepterait mieux d'acquiescer à des demandes présentées par des personnes pour lesquelles on a de l'estime.

Nous sommes intéressés à savoir dans quelle mesure les intervieweurs utilisent ces principes pour persuader les personnes faisant partie d'un échantillon de participer à l'enquête correspondante.

On soutient que les intervieweurs qui font une utilisation appropriée de chacune de ces stratégies auront vraisemblablement un plus grand succès pour ce qui est de persuader les personnes, dans l'échantillon, qui sont peu disposées à collaborer à l'enquête à y participer. Toutefois, l'utilisation, sans distinction, de telles techniques de persuasion dans toutes les situations peut se retourner contre l'intervieweur. Par exemple, si l'on fait appel au principe de l'autorité dans des régions où il y a beaucoup de méfiance à l'endroit du gouvernement, cela peut fort bien avoir un effet négatif sur la collaboration. L'utilisation de ces principes d'acquiescement peut ne pas être universellement efficace dans toutes les situations ou pour toutes les personnes faisant partie d'un échantillon.

Ainsi, il ne suffit pas de savoir si ces techniques de persuasion sont utilisées par les intervieweurs, mais aussi **comment** elles le sont. Deux concepts nous intéressent ici. L'un d'entre eux porte sur le nombre de techniques différentes dont un intervieweur dispose et le second a trait à la mesure dans laquelle on applique ces techniques de façon appropriée. Nous désignons le premier de ces concepts par l'expression "répertoire de techniques" dont l'intervieweur dispose. Un intervieweur débutant peut apprendre une ou deux présentations "toutes faites" pendant sa formation puis les utiliser avec toutes les personnes faisant partie d'un échantillon avec lesquelles il a travaillé. Par contre, l'intervieweur expérimenté possède un répertoire étendu de présentations parmi lesquelles il peut choisir et qu'il peut appliquer en fonction de la situation.

Le second concept est celui de l'application appropriée des aptitudes ou des techniques dont l'intervieweur dispose. Nous désignons ce concept par l'expression "adaptation". On s'attend à ce qu'un intervieweur soit un "diagnostiqueur psychologique avisé" (Cannell 1964), qu'il soit en mesure d'évaluer une situation rapidement et d'appliquer les messages persuasifs appropriés. Ces aptitudes s'acquièrent par l'expérience, soit au travail, soit dans la vie en général. Il se peut que l'intervieweur débute avec moins d'aptitudes et moins de confiance, adhère, de façon rigide, à un petit nombre de méthodes "qui ont fait leurs preuves". L'intervieweur expérimenté est plus en mesure d'adapter sa méthode à chaque répondant éventuel. Il se peut que la faculté d'adaptation ainsi que l'application appropriée de techniques de persuasion soient plus critiques que les techniques ou les comportements réels eux-mêmes. Si c'est le cas, il devrait être possible d'élaborer un modèle parcomique qui n'utilise que les derniers concepts et d'éliminer les comportements particuliers mesurés.

2.4 Zone d'affectation

Afin d'examiner l'effet des intervieweurs sur la participation à une enquête, nous devons tenir compte du fait qu'on leur confie différentes zones à interviewer. Idéalement, dans le plan d'expérience on affecterait de façon aléatoire les intervieweurs à des régions de l'échantillon, supprimant toute confusion statistique entre les caractéristiques de l'intervieweur et celles de la population. Sans une telle randomisation, nous tentons de préciser les caractéristiques de la population qui ont de l'importance pour le taux de réponse et de les "neutraliser" statistiquement.

à l'aise pour s'occuper de la grande variété de personnes qui font partie d'un échantillon avec lesquelles ils peuvent avoir à traiter et de zones d'affection qui peuvent leur être confiées. Après ce point, il se peut que les années additionnelles d'expérience ne produisent pas d'autres gains dans les taux de réponse.

Une autre hypothèse est que l'auto-sélection plutôt que l'expérience produit des taux de réponse plus élevés parmi les intervieweurs qui ont été en fonction plus longtemps. En d'autres mots, ce n'est pas le fait que les intervieweurs s'améliorent avec le temps, mais que les meilleurs intervieweurs ont tendance à rester alors que les intervieweurs moins bons cessent de faire ce travail. Nous croyons qu'une combinaison de ces deux facteurs explique les variations dans le rendement des intervieweurs. Toutefois, on ne peut vérifier l'hypothèse de l'auto-sélection dans une enquête transversale comme celle dont nous parlons et il faut donc faire preuve de prudence quand on tire des conclusions à partir de ces analyses.

Si les intervieweurs expérimentés obtiennent des taux de réponse plus élevés, nous supposons que cela se produit suite aux effets combinés des attentes des intervieweurs (*p. ex.*, la confiance) et de leur comportement (*p. ex.*, une présentation orale efficace). Il faut remarquer que nous ne supposons aucun effet direct de l'expérience sur les taux de réponse. En d'autres mots, est-il possible de déterminer les aptitudes et les comportements des intervieweurs qui peuvent expliquer les différences possibles dans les taux de réponse?

2.2 Attentes de l'intervieweur

On fait l'hypothèse que des attentes positives de la part des intervieweurs contribuent à des taux de réponse plus élevés. Il est probable que les intervieweurs qui ont une plus grande confiance dans leur aptitude à persuader les personnes faisant partie d'un échantillon donné à participer à l'enquête correspondent, qui croient en la légitimité du travail qu'ils effectuent et qui sont confiants que la majorité des personnes acceptent de participer aux enquêtes obtiendront des taux de réponse plus élevés que ceux qui n'ont pas cette attitude. L'étude de Singer, Frankel et Glassman (1983), dans laquelle on a trouvé que les intervieweurs qui prévoyaient, avant l'enquête, que la tâche de persuader les répondants était 'assez facile', ont obtenu des taux de réponse plus élevés que ceux qui croyaient que la tâche était 'assez difficile', apporte un certain appui empirique à cet argument.

2.3 Comportement de l'intervieweur

Pour ce qui est des comportements des intervieweurs, nous cherchons à déterminer les mécanismes qui permettent de transformer une expérience plus considérable et des attentes positives en des taux de réponse plus élevés. On peut comparer le comportement des intervieweurs pour ce qui est d'obtenir la collaboration des personnes faisant partie de l'échantillon à celui d'autres 'professionnels de l'acquisition' (comme les vendeurs, les collecteurs de fonds, etc.). Après un examen approfondi des preuves basées tant sur des expériences que sur l'observation, Cialdini (1984, 1990) détermine six principes d'acquisition utilisés pour décider si l'on va accéder à une demande. Voici, brièvement, ces principes:

- a) Échange: une personne accepterait mieux d'acquiescer à une demande dans la mesure où l'acquiescement constitue un témoignage de reconnaissance pour ce qu'on a perçu comme un cadeau, une faveur ou une concession.
- b) Cohérence: Après s'être engagée dans une position, une personne accepterait mieux d'acquiescer à des demandes relatives à des comportements qui sont cohérents avec cette position.
- c) Validation sociale: On accepterait mieux d'acquiescer à une demande dans la mesure où l'on croit que des personnes qui nous ressemblent acquiesceraient à cette demande.

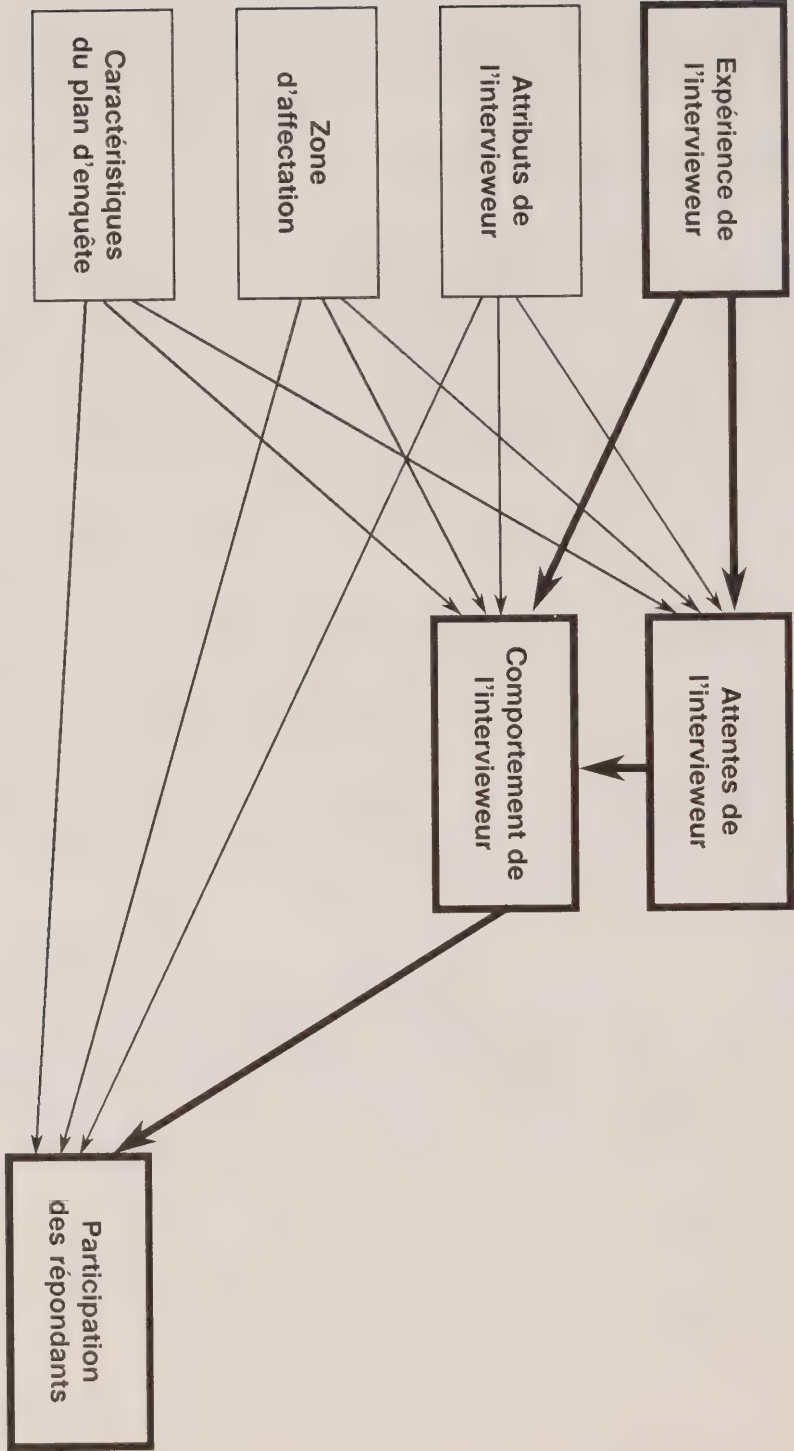


Figure 1. Modèle de participation à une enquête, Rôle de l'intervieweur.

Dans le présent article, nous examinons le rôle de diverses caractéristiques des intervieweurs, particulièrement l'expérience, pour obtenir la collaboration des répondants. Il faut remarquer que l'intervieur ne représente qu'un élément d'un ensemble important de facteurs qui peuvent avoir un effet sur la participation à une enquête. Ces facteurs comprennent des caractéristiques des répondants, l'interaction entre le répondant et l'intervieur, des caractéristiques du plan d'enquête ainsi que des facteurs contextuels et liés à la situation. Pour une étude de ces facteurs, se reporter à Groves, Cialdini et Couper (1992).

Nous devons aussi remarquer qu'il se peut que des modèles différents soient plus appropriés pour des composantes différentes de la non-réponse. Par exemple, la motivation des intervieweurs, leur ténacité et l'effort qu'ils consacrent peuvent être plus importants pour réduire le nombre de cas où l'on n'a pu entrer en communication avec le répondant, alors que les aptitudes liées à la persuasion jouent un plus grand rôle dans la partie refus de la non-réponse. Les données analysées ici ne nous permettent pas de faire la distinction entre ces composantes de la non-réponse. Conséquemment, les modèles testés n'ont pas tout le pouvoir explicatif qu'ils pourraient posséder.

Dans le présent article, nous nous intéresserons à deux questions: a) les intervieweurs expérimentés obtiennent-ils des taux de réponse plus élevés? b) dans l'affirmative, quels sont les mécanismes à la base du rapport entre l'expérience et les taux? Ces questions sont importantes pour l'ensemble des chercheurs s'intéressant aux enquêtes. Si l'on peut faire adopter aux intervieweurs débutants les comportements des intervieweurs expérimentés qui obtiennent des taux de réponse élevés, alors il se pourrait que les premiers connaissent le même succès que ces derniers. Si ce n'est pas le cas, alors l'importance de réduire le taux de roulement parmi les intervieweurs expérimentés demeure élevée pour les organismes qui réalisent des enquêtes.

2. TRAVAUX EN VUE D'OBTENIR UN MODÈLE DE PARTICIPATION À UNE ENQUÊTE

On peut déterminer un certain nombre de caractéristiques des intervieweurs qui peuvent avoir une incidence sur la participation à une enquête. Ces caractéristiques sont illustrées à la figure 1. Nous étudierons les effets de l'expérience de l'intervieur, de ses attentes et de son comportement sur les taux de réponse, en "neutralisant" les caractéristiques de la région d'affectation ainsi que du plan d'enquête. Nous examinerons, à tour de rôle, chacun des ensembles de variables.

2.1 Expérience de l'intervieur

Tout d'abord, on s'attend à ce que l'expérience des intervieweurs ait un effet positif sur les taux de réponse que ces derniers obtiennent. Cela découle des leçons apprises par tâtonnement lors de l'application de différentes techniques au fil des ans et de diverses lignes directrices relatives à la formation ainsi que des expériences acquises lors de diverses enquêtes. L'expérience a donc deux composantes: la durée et l'étendue. Le nombre d'années pendant lesquelles une personne a travaillé comme intervieweur pourrait représenter la durée de l'expérience. Un indicateur de l'étendue de l'expérience est le nombre d'organismes différents pour lesquels un intervieweur a travaillé ou le nombre de genres d'études différentes auxquelles un intervieweur a travaillé. On soutient que la durée et l'étendue de l'expérience servent toutes deux à accroître la variété des différentes situations d'interview auxquelles un intervieweur est exposé.

Nous nous attendons à ce que le rapport entre la durée de l'expérience (telle que mesurée par la durée des fonctions) et les taux de réponse soit curviligne. L'expérience acquise au cours des premières années de réalisation d'interviews aura une plus grande incidence sur les taux de réponse que celle qui est acquise plus tard. Après un certain point, le nombre de nouvelles situations auxquelles un intervieweur a à faire face diminue et les intervieweurs deviennent plus

Le rôle de l'intervieur dans la participation aux enquêtes

MICK P. COUPER et ROBERT M. GROVES¹

RÉSUMÉ

À l'aide de données tirées d'une enquête portant sur des intervieweurs du U.S. Census Bureau, cet article étudie si des intervieweurs expérimentés obtiennent des taux de réponse plus élevés que des intervieweurs inexpérimentés, en "neutralisant" les différences dans le plan d'enquête et dans les attributs des populations affectées aux intervieweurs. Après avoir démontré que le rapport est positif et curviligne, l'étude tente d'expliquer les mécanismes qui permettent aux intervieweurs expérimentés d'obtenir ces taux de réponse et élabore sur la nature du rapport. L'article examine quels comportements et quelles attitudes sont à la base du plus grand succès dans l'espoir que ces éléments pourraient être inculqués aux stagiaires.

MOTS CLÉS : Intervieweurs; non-réponse; taux de réponse; participation à une enquête.

1. INTRODUCTION

Les spécialistes des méthodes d'enquête soupçonnent depuis longtemps que l'intervieur est une source importante de variation dans les taux de réponse. Les indicateurs de cette situation comprennent les différences observées parmi les stagiaires dans l'aptitude à observer et à appliquer les lignes directrices relatives aux interviewers, une variation dans le taux de données manquantes pour les questions entre les interviewers, les taux de réponse d'interviewers particuliers ainsi que l'aptitude de certains interviewers à obtenir des réponses auprès d'enquêtés qui avaient refusé de collaborer avec d'autres interviewers. Toutefois, le fait que les interviewers travaillent souvent dans des sous-populations différentes et qu'ils font donc face à des défis différents pour remplir leurs tâches a une influence sur plusieurs de ces indicateurs.

Une bonne partie de ce que nous croyons à propos de l'incidence de l'intervieur sur la participation à une enquête n'a pas encore été vérifiée ou demeure peu concluante. Dans une étude souvent citée, Durbin et Stuart (1951) ont trouvé que les interviewers expérimentés étaient incontestablement supérieurs à des volontaires étudiants pour ce qui est des taux de réponse. Groves et Fultz (1985) ont trouvé que les interviewers débutants (qui n'étaient entrés en fonction que depuis 1 à 6 mois) avaient les taux de refus les plus élevés au cours d'une enquête téléphonique. Dans une étude citée par Inderfurth (1972), les taux de non-réponse pour les interviewers du Census Bureau formés en 1962 et en 1963 ont diminué régulièrement au cours des premiers mois d'emploi pour atteindre le niveau des interviewers expérimentés après 22 mois. Par contre, Singer, Frankel et Glassman (1983, p. 74) ont trouvé que l'effet de l'expérience sur les taux de réponse au cours d'une enquête téléphonique était contraire à l'intuition, c'est-à-dire, que les interviewers plus expérimentés n'ont pas obtenu des taux de réponse plus élevés. Ils l'ont remarquer, cependant, que ce résultat est basé sur seulement six interviewers. Dans une étude portant sur 16 interviewers travaillant sur le terrain en Suède, Schybergger (1967) a trouvé que les taux de non-réponse étaient **plus élevés** pour les interviewers expérimentés que pour les interviewers qui avaient été recrutés peu de temps auparavant. En bref, les résultats empiriques n'appuient pas toujours la croyance universelle que les interviewers expérimentés ont plus de succès.

¹ Mick P. Couper et Robert M. Groves, U.S. Bureau of the Census et University of Michigan. Bureau 2315-3, Bureau of the Census, Washington, DC 20233.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à Pierre Lavallée, qui, le premier, a établi les expressions relatives aux fractions de sondage de seconde phase conditionnellement optimales utilisées dans la méthode approximative. Ils remercient également l'arbitre et le rédacteur associé qui leur ont formulé des commentaires utiles. C.F.J. Wu est bénéficiaire d'une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

ARMSTRONG, J.B., BLOCK, C., et SRINATH, K.P. (1991). Two-phase sampling of tax records for business surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 228-233.

BOOTH, G., et SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.

CHOUDHRY, G.H., LAVALLÉE, P., et HIDIROGLOU, M. (1989a). Two-phase sample design for tax data. Document non publié, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

CHOUDHRY, G.H., LAVALLÉE, P., et HIDIROGLOU, M. (1989b). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3ième éd.). New York: John Wiley.

IMSL (1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

SCHITTOKOWSKI, K. (1985). NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition? *Journal of the American Statistical Association*, 84, 916-921.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.

Tableau 2

Résultats pour la méthode exacte et la méthode approximative

Méthode	Exacte - Valeur de départ			Coût total (\$)	Nombre de cas pour lesquelles la méthode a produit le meilleur résultat*
	I	II	III		
Approximative	Valeur de départ				
	I	II	III		
					Nombre de cas pour lesquelles la méthode a produit le meilleur résultat*
					48
					17
					1
					122779
					139347
					200998
					130228

* Pour deux cas, les valeurs de départ I et II ont produit le même résultat et celui-ci est supérieur au résultat obtenu au moyen de la valeur de départ III. C'est pourquoi la somme des chiffres de cette ligne est 66 plutôt que 64.

fonction de coût entre deux itérations était inférieure à 10⁻⁴. La version d'IMSL de l'algorithme de programmation quadratique successive de Schittkowski était utilisée pour résoudre les problèmes de programmation non linéaire.

Les résultats de l'étude empirique figurent dans le tableau 2. On y trouve le coût total de l'échantillonnage calculé selon quatre solutions, de même que le nombre de cas "CT12" pour lesquelles chacune des valeurs de départ utilisées pour la méthode exacte a produit de meilleurs résultats que d'autres valeurs de départ. Nous n'avons pas indiqué les coûts du calcul puisqu'ils n'étaient pas suffisamment élevés pour avoir une influence quelconque.

Les résultats montrent que la méthode approximative offre les meilleurs valeurs de départ pour la méthode exacte. Bien que la valeur de départ II ait donné de meilleurs résultats que la valeur de départ I pour 17 cas "CT12", le coût total associé à la valeur II est plus élevé que celui établi à l'aide de la méthode approximative. La méthode exacte laisse à désirer lorsque les valeurs de départ correspondent à un ensemble aléatoire de fractions de sondage de première phase possibles.

Bien que le coût total établi au moyen de la valeur de départ I de la méthode exacte n'est que 5,7% moins élevé que le coût établi à l'aide de la méthode approximative, il convient de souligner que la méthode exacte (avec valeur de départ I) sera toujours supérieure à la méthode approximative. Celle-ci a donné de meilleurs résultats que celle-ci pour 42 cas.

5. CONCLUSION

Dans les deux premières sections de cet article, nous avons exposé le problème de la répartition d'un échantillon pour les plans de sondage à deux phases comme un problème d'optimisation sous contrainte. Si le nombre de variables et de contraintes est peu élevé, on peut obtenir une solution par l'application directe de méthodes numériques. En revanche, l'approche directe laisse à désirer lorsque les variables et les contraintes sont en grand nombre.

En exploitant la structure mathématique du problème, on peut reformuler celui-ci en deux sous-problèmes: premièrement, un problème de programmation convexe avec contraintes linéaires qui renferme beaucoup moins de variables et deuxièmement, un problème que l'on peut résoudre sans recourir à des méthodes numériques. L'algorithme que nous avons proposé dans la section 2 consiste en des itérations entre les deux sous-problèmes. Plus simple sur le plan du calcul, il est plus efficace que l'approche directe pour résoudre des problèmes qui renferment un grand nombre de variables et de contraintes. Dans la section 3, nous avons présenté une méthode approximative de répartition de l'échantillon qui ne nécessite pas l'emploi de méthodes numériques. L'étude empirique de la section 4 montre que la méthode approximative offre une valeur de départ particulièrement bonne pour l'algorithme proposé dans la section 2.

Notons qu'une répartition ne sera possible que si D^{T_h} est positif. En substituant (13) dans (12), on obtient

$$v_{gh}^* = (A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*) \cdot \sum_{g \in T_h} (N_{gh} \cdot A_{gh})^{1/2} / D^{T_h}. \tag{14}$$

Si v_{gh}^* est supérieure à un pour certaines strates de seconde phase g_h , on peut bien sûr utiliser la méthode de "surrépartition" mentionnée plus haut. Notons que l'équation (14) produit aussi la solution voulue pour l'étape (ii) de chaque itération de la méthode exacte.

4. ETUDE EMPIRIQUE

La méthode approximative est utile à deux points de vue. Premièrement, elle constitue une bonne valeur de départ pour la méthode exacte et deuxièmement, elle est probablement plus facile à appliquer dans la pratique. Dans cette section, nous présentons les résultats d'une étude comparative empirique dans laquelle nous avons utilisé des données de la province de Québec pour l'année d'imposition 1988. Nous communiquons les résultats obtenus au moyen de la méthode exacte, pour diverses valeurs de départ, comme au moyen de la méthode approximative. Etant donné que les quantités N_{gh}^* , Y_h et S_{gh}^* , nécessaires aux deux méthodes, étaient inconnues, nous les avons estimées à l'aide des données.

Nous avons procédé à une stratification selon la taille, comme dans la vraie enquête: quatre strates à tirage partiel et une à tirage complet. Des schèmes de répartition ont été déterminés pour 64 cases, "CT12" (toutes tirées des données du Québec sauf quelques-unes à effectif réduit). Le nombre de fractions de sondage calculé au cours de cette opération variait de 8 à 92, avec une médiane de 24. Le nombre de contraintes, lui, variait de 9 à 15, avec une médiane de 31. Une vingtaine de cases, "CT12" renfermaient plus de 35 variables et dix-huit de ces cases renfermaient aussi plus de 50 contraintes. En tout, nous avions 1,850 strates de seconde phase, qui comptaient environ 230,000 unités de population.

Le coût de l'échantillonnage à la première phase a été établi à \$1.40 l'unité; ce coût comprenait le microfilmage ou la reproduction par photocopie des déclarations d'impôt à Revenu Canada, l'envoi de l'information à Statistique Canada et la détermination des codes à quatre chiffres de la Classification type des industries. Quant au coût de l'échantillonnage de la seconde phase, qui correspondait essentiellement au coût de la transcription de valeurs pour des variables financières, il a été établi à \$7.00. Ces coûts reflètent assez bien ceux enregistrés normalement dans la vraie enquête.

Pour calculer les schèmes de répartition à l'aide de la méthode exacte, nous nous sommes servis de trois valeurs de départ: I – la solution de la méthode approximative; II – l'ensemble des fractions de sondage de la première phase posées égales à un, plus les fractions de la seconde phase conditionnellement optimales correspondantes; et III – un ensemble aléatoire de fractions de sondage de première phase possibles, avec les fractions de seconde phase conditionnellement optimales correspondantes. En outre, nous avons éprouvé la méthode exacte avec une perturbation de la fraction de sondage de première phase pour la strate g était $v_{g(0)}^* = 0.1 + 0.9 \cdot v_g^*$, où v_g^* est la solution de la méthode approximative. Les fractions de sondage de la seconde phase ont pris au départ des valeurs optimales, étant donné la valeur perturbée des fractions de la première phase. La valeur de départ III a été perturbée de façon analogue. En ce qui a trait à la valeur de départ II, la valeur perturbée de la fraction de sondage était $v_{g(0)}^* = 0.1 + 0.9 \cdot v_g^*$, où v_g^* est optimale, à condition qu'il y ait recensement à la première phase de l'échantillonnage. Pour chaque valeur de départ, nous avons retenu le meilleur résultat observé, peu importe qu'il ait été obtenu à l'aide de la valeur autheutique ou de la valeur perturbée correspondante. La convergence était constatée dès que la variation relative de la

les fractions de sondage de la seconde phase égales à un pour les strates g qui ne sont pas visées par l'opération. Normalement, on devrait avoir l'ensemble $\Gamma = \{1, 2, \dots, G\}$, mais à cause du phénomène de "surépartition" qui survient durant la répartition de l'échantillon de la seconde phase, par exemple, Γ peut ne pas contenir tous les entiers de 1 à G . Répartir l'échan-
tillon de la seconde phase équivalent à déterminer le minimum de

$$(9) \qquad F^h = \sum_{g \in \Gamma} v_{gh} \cdot v_g^* \cdot N_{gh}$$

par rapport à v_{gh} , $g \in \Gamma$, étant donné les contraintes

$$(10) \qquad \sum_{g \in \Gamma} \left(\frac{1}{v_{gh}} - 1 \right) \cdot \frac{A_{gh}}{v_g^*} \leq M_h,$$

$$(11) \qquad 0 < v_{gh} \leq 1, \quad g \in \Gamma,$$

où

$$M_h = C_h^2 \cdot Y_h^2 \cdot \sum_g \left(\frac{1}{v_g^*} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Notons que le nombre prévu d'unités appartenant au domaine h (CT14) dans l'échantillon de la seconde phase pour la strate g , $v_{gh}^* \cdot N_{gh}$, figure dans l'équation (9). On peut montrer facilement que (9) prend une valeur minimum lorsque l'égalité dans l'équation (10) est satisfaite. Par conséquent, la minimisation équivalent à déterminer le point critique du lagrangien

$$L^h = \sum_{g \in \Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} + \lambda \cdot \left(M_h - \sum_{g \in \Gamma} \left(\frac{1}{v_{gh}} - 1 \right) \cdot \frac{A_{gh}}{v_g^*} \right),$$

par rapport à v_{gh} , $g \in \Gamma$, et à λ , étant donné les contraintes

$$0 < v_{gh} \leq 1, \quad g \in \Gamma.$$

En posant les dérivées premières de L^h égales à zéro et en simplifiant, on obtient

$$(12) \qquad v_{gh} = (-\lambda \cdot A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*), \quad g \in \Gamma,$$

$$(13) \qquad (-\lambda)^{1/2} \sum_g (N_{gh} \cdot A_{gh})^{1/2} / D^h,$$

où

$$D^h = C_h^2 \cdot Y_h^2 \cdot \sum_{g \in \Gamma} \left(\frac{1}{v_g^*} \right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^*} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

par rapport à $v_{g|h}^*$, $g = 1, 2, \dots, G$. Le symbole $v_{g|h}^*$ indique que la détermination de la fraction de sondage pour la strate g est assujettie à une seule contrainte de précision, à savoir celle relative au domaine h (CTI4), h étant fixe. En particulier, la minimisation doit s'effectuer en égard aux contraintes

$$(4) \quad \sum_{g=1}^G \left(\frac{1}{v_{g|h}^*} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2,$$

$$(5) \quad 0 < v_{g|h} \leq 1, \quad g = 1, 2, \dots, G.$$

On peut montrer que le minimum de (3) est obtenu lorsque l'égalité dans l'équation (4) est satisfaite, de sorte que le problème défini par les équations (3), (4) et (5) équivaut à calculer le point critique du lagrangien

$$L = \sum_{g=1}^G v_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_{g=1}^G \left(\frac{1}{v_{g|h}^*} - 1 \right) \cdot (A_{gh} + B_{gh}) \right].$$

En posant les dérivées par rapport à $v_{g|h}$ égales à zéro, on obtient

$$(6) \quad v_{g|h} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot (-\lambda)^{1/2}, \quad g = 1, 2, \dots, G.$$

En posant $\partial L / \partial \lambda = 0$, on obtient

$$(7) \quad (-\lambda)^{1/2} = \sum_{g=1}^G ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \cdot \left(C_h^2 \cdot Y_h^2 + \sum_{g=1}^G (A_{gh} + B_{gh}) \right).$$

Après avoir substitué (7) dans (6), on obtient la fraction de sondage optimale pour la strate g , étant donné une seule contrainte de précision, celle relative au domaine h (CTI4),

$$v_{g|h}^* = ((A_{gh} + B_{gh})/N_g)^{1/2}.$$

$$(8) \quad \sum_{g=1}^G ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \cdot \left(C_h^2 \cdot Y_h^2 + \sum_{g=1}^G (A_{gh} + B_{gh}) \right).$$

Si une ou plusieurs des fractions de sondage définies par l'expression (8) sont supérieures à un, on peut poser ces fractions égales à un et procéder à une répartition avec un nombre moindre de strates. On reconnaît dans cette technique la méthode de "surrépartition" dont fait état Cochran (1977). Il est nécessaire de calculer (8) pour $h = 1, 2, \dots, H$. La fraction de sondage approximative de la première phase pour la strate g , $v_{g|*}^*$, est posée égale à la valeur la plus élevée de l'ensemble $\{v_{g|h}^*, h = 1, 2, \dots, H\}$ pour $g = 1, 2, \dots, G$; de cette manière, nous sommes sûrs que la contrainte de précision relative à chaque domaine "CTI4" sera satisfaite.

Étant donné les fractions de sondage de la première phase, il est facile de calculer les fractions de sondage optimales de la seconde phase. Supposons que, pour la case h formée par le croisement des codes à deux chiffres de la CTI (CTI2) et des provinces, les strates g visées par l'opération de répartition correspondent à un ensemble de nombres entiers, T . Nous posons

$$X_{(i)}^g \geq 0, \quad g = 1, 2, \dots, G.$$

(ii) calcul de $v_{(i)}^g = 1/(X_{(i)}^g + 1)$, $g = 1, 2, \dots, G$. Minimisation, pour chaque valeur de $h = 1, 2, \dots, H$, prise individuellement, de

$$F_h = \sum_{g=1}^G v_{(i)}^g \cdot v_{gh}^{(i)} \cdot N_{gh}$$

par rapport à $v_{gh}^{(i)}$, $g = 1, 2, \dots, G$, étant donné les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_{g=1}^G \left(\frac{v_{(i)}^g}{1 \cdot v_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_{g=1}^G \left(\frac{v_{(i)}^g}{1} - 1 \right) \cdot B_{gh} \geq 0, \\ 0 < v_{(i)}^g \leq 1, \quad g = 1, 2, \dots, G,$$

où h est considéré comme fixe.

Nous verrons dans la section 3 qu'il n'est pas nécessaire de recourir à des méthodes numériques pour résoudre l'étape (ii). La méthode exacte ne nécessite donc que la résolution d'une série de problèmes de programmation convexe ne renfermant chacun que G variables. Il est beaucoup plus facile de résoudre un problème de programmation convexe qu'un problème de programmation non linéaire général. Dans le premier cas, une solution locale est aussi une solution globale.

Posons $F_{(i)}$ comme la valeur de la fonction de coût, (i), calculée au moyen de $v_{(i)}^g$ et de $v_{gh}^{(i)}$. Les valeurs $F_{(i)}$ forment une suite monotonement décroissante et, de ce fait, tendent vers une limite. C'est la valeur de départ qui détermine si cette limite et les fractions de sondage correspondantes conduisent au minimum global, cela à cause de la géométrie des contraintes définies en (2). Dans la pratique, il faudrait tester plusieurs valeurs de départ en vue de trouver la meilleure solution. Une de ces valeurs nous est fournie par la "méthode approximative", qui ne nécessite pas d'itérations. Cette méthode est décrite dans la section suivante.

3. MÉTHODE APPROXIMATIVE

Dans cette section, nous présentons une méthode de répartition qui se rapproche de la répartition optimale. Elle a été proposée la première fois par Choudhry, Lavallée et Hidiroglou (1989a). Après avoir supposé que toutes les fractions de sondage de la seconde phase sont égales à un, on détermine une formule approchée de répartition optimale pour l'échantillon de la première phase. On répartit ensuite l'échantillon de la seconde phase, compte tenu des fractions de sondage de la première phase. Comme le coût de l'échantillonnage d'une unité dans l'autre phase ne dépend pas de la strate dans laquelle se trouve l'unité, minimiser le coût revient à minimiser la taille de l'échantillon à chaque étape de la méthode.

La première étape de la méthode approximative consiste à trouver une solution approchée au problème de la répartition optimale pour un plan d'échantillonnage à une phase. Pour cela, on doit calculer le minimum – pour chaque valeur de h prise individuellement – de

$$F_{(h)} = \sum_{g=1}^G v_{gh}^{(h)} \cdot N_g \tag{3}$$

Tableau 1

Résultats pour la méthode directe et la méthode exacte

CTI 2	Nombre de variables	Nombre de contraintes	Coût (\$) - méthode directe	Coût (\$) - méthode exacte
30	62	86	5155**	1897
35	37	51	551	512
39	38	50	1667	1450
427*	39	48	27528**	3383

* En ce qui a trait aux industries de la construction, on se sert des codes à trois chiffres pour la stratification de la première phase.
** Le programme d'IMSLS est interrompu à cause d'une erreur interne, qui n'a pu être corrigée après consultation de la documentation.

calculé au moyen de la méthode que nous appelons "méthode exacte" et qui est décrite plus bas. Les données du tableau indiquent que l'application directe de l'algorithme de Schittkowski à la manière d'IMSLS est inadéquate pour des cas "CTI2" lorsqu'il y a un grand nombre de variables et de contraintes.

La méthode exacte repose sur une simplification majeure du problème défini par (1) et (2); précisément, nous reformulons le problème en deux grandes étapes, chacune d'elles pouvant recevoir une solution itérative. Dans la première étape, on minimise (1) par rapport à v_g , $g = 1, 2, \dots, G$, étant donné la valeur de toutes les fractions de sondage de la seconde phase. Cette étape nécessite l'utilisation de techniques d'optimisation non linéaire. La deuxième étape consiste à minimiser (1) par rapport aux fractions de sondage de la seconde phase, étant donné les fractions de sondage de la première phase calculées dans la première étape. Cette minimisation ne nécessite pas d'itérations puisque sa solution a une forme analytique fermée. En outre, elle peut se faire pour chaque valeur de h prise individuellement, $h = 1, 2, \dots, H$. Une fois la seconde étape terminée, on répète la première étape et on poursuit le processus itératif. On constate la convergence lorsque la variation de la fonction de coût entre deux itérations devient négligeable.

Posons $v_g^{(i)}$ et $v_{gh}^{(i)}$ comme les estimations des valeurs optimales de v_g et de v_{gh} obtenues après i itérations (chaque itération comprenant une double exécution des deux étapes décrites ci-dessus). Au début de l'itération $i + 1$, il faut transformer les variables définies par l'expression $X_{(i+1)}^g = 1/v_{(i+1)}^g - 1$. Cette opération a pour but de redéfinir le problème d'optimisation qui se pose dans la première étape de l'itération comme un problème renfermant des contraintes linéaires et une fonction-objectif convexe, c'est-à-dire comme un problème de programmation convexe. Les problèmes de ce genre sont plus faciles à résoudre.

Plus précisément, chaque itération comporte les opérations suivantes:

(i) minimisation de

$$F = \sum^g \left(N_g + \frac{K_1}{K_2} \sum^h v_{(l-1)}^{gh} \cdot N_{gh} \right) / \left(X_{(i)}^g + 1 \right)$$

par rapport à $X_{(i)}^g$, $g = 1, 2, \dots, G$, étant donné les contraintes

$$C_2^h \cdot Y_2^h - \sum^g \left(X_{(i)}^g + 1 \cdot \frac{v_{(l-1)}^{gh}}{u_{(l-1)}^{gh}} - 1 \right) \cdot A_{gh} - \sum^g X_{(i)}^g \cdot B_{gh} \geq 0, \quad h = 1, 2, \dots, H$$

2. MÉTHODE EXACTE

Dans cette section, nous exposons le problème de la répartition optimale et nous lui opposons une solution itérative appelée "méthode exacte". Pour exposer ce problème par rapport à l'échantillonnage à deux phases de dossiers fiscaux, il suffit de considérer une case formée par le croisement des codes à deux chiffres de la CTI (CTI2) et d'une province en particulier qui contient N unités. Le coût de l'échantillonnage d'une unité dans la première phase est K_1 , quelle que soit la strate dans laquelle se trouve cette unité, tandis qu'il est de K_2 dans la seconde phase, peu importe la strate. Suivant l'échantillonnage de Bernoulli, la fonction de coût est

$$F^* = K_1 \cdot \sum_g n_{gh}' + K_2 \cdot \sum_h \sum_g n_{gh}.$$

Puisque la taille des échantillons, n_g' et n_{gh} , est aléatoire, nous utilisons l'espérance de coût

$$(1) \quad F = K_1 \cdot \sum_g v_g \cdot N_g + K_2 \cdot \sum_h \sum_g v_g \cdot v_{gh} \cdot N_{gh}.$$

Rao (1973) et Smith (1989) se servent aussi de l'espérance mathématique de fonctions de coût aléatoires pour résoudre des problèmes de répartition dans des plans à deux phases. Pour ce qui a trait à l'échantillonnage de dossiers fiscaux, le coût total pour une province équivalant à la somme des coûts de l'échantillonnage pour toutes les cases "CTI2" dans la province. Le coefficient de variation estimé du coût de l'échantillonnage à deux phases de dossiers fiscaux pour la province de Québec, fondé sur des données de 1988, était environ 1,85%. Ce coefficient était moindre pour le coût de l'échantillonnage à l'échelle nationale.

Il est nécessaire de minimiser (1) par rapport à v_g , $g = 1, 2, \dots, G$, et à v_{gh} , $g = 1, 2, \dots, G$, $h = 1, 2, \dots, H$, étant donné les contraintes

$$\sum_g \left(\frac{1}{1 + v_g \cdot v_{gh}} \cdot A_{gh} + \sum_g \left(\frac{1}{1 + v_g} - 1 \right) \cdot B_{gh} \leq C_h^2 \cdot Y_h^2, \quad h = 1, 2, \dots, H, \quad (2) \right. \\ \left. 0 < v_g \leq 1, \quad g = 1, 2, \dots, G, \right. \\ \left. 0 < v_{gh} \leq 1, \quad g = 1, 2, \dots, G, \quad h = 1, 2, \dots, H, \right.$$

où C_h désigne le coefficient de variation cible pour le domaine h (CTI4).

Lorsqu'on a tenté de résoudre directement le problème en appliquant l'algorithme de programmation quadratique successive de Schittkowski (1985) à la manière d'IMSL (1987), on a obtenu des résultats variables. L'algorithme fonctionnait bien pour des problèmes qui renfermaient un petit nombre de variables et de contraintes. Cependant, il n'a pas été possible d'obtenir des solutions satisfaisantes pour les problèmes où il y avait plus de 35 variables ou plus de 50 contraintes environ.

Le tableau 1 contient quelques chiffres de coût obtenus par l'application directe de l'algorithme de Schittkowski en ce qui concerne l'échantillonnage de dossiers fiscaux. Cette opération visait à résoudre les problèmes de répartition liés à certaines cases "CTI2" dans la province de Québec où il fallait composer avec un grand nombre de variables et de contraintes; à cette fin, nous nous sommes servis de données pour l'année d'imposition 1988. Avec l'approche directe, toutes les fractions de sondage (première phase comme seconde phase) étaient posées égales à un au départ. Le tableau 1 donne aussi le coût (moins élevé que l'autre)

Statistique Canada se procure des renseignements sur la population des déclarants auprès des Revenu Canada. L'organisme statistique est tenu de produire des estimations de variables financières pour des domaines définis par le croisement des provinces et des codes à quatre chiffres de la Classification type des industries (CTI4). Or, à Revenu Canada, seuls les codes à deux chiffres offrent un niveau de précision acceptable. Afin de normaliser le niveau de précision des estimations pour les domaines définis par le croisement des 'CTI4' et des provinces, on a mis en application un plan d'échantillonnage à deux phases. Dans la première phase, on prélève un échantillon de déclarants à Revenu Canada dans des strates définies en fonction du code à deux chiffres de la CTI et du revenu brut de l'entrepris (taille). Avant de passer à la seconde phase de l'échantillonnage, Statistique Canada attribue un code à quatre chiffres de la CTI à chaque unité échantillonnée; ce code est réputé plus précis que ceux fournis par Revenu Canada. Dans la seconde phase de l'échantillonnage, on se sert de strates définies en fonction du code à quatre chiffres de la CTI et de la taille de l'entrepris. Les limites de classe sont les mêmes dans les deux phases. Pour une description détaillée du plan d'échantillonnage, le lecteur est prié de se référer à Choudhry, Lavallée et Hidiroglou (1989b).

L'échantillon de la première phase est formé suivant un échantillonnage de Bernoulli (appelé aussi échantillonnage de Poisson). Supposons que le déclarant i d'une case particulière formée par le croisement des provinces et des codes à deux chiffres de la CTI (CTI2) appartient à la strate de première phase g . Pour déterminer si ce déclarant appartient à l'échantillon de première phase, on génère un nombre pseudo-aléatoire compris dans l'intervalle $(0,1)$, disons R_i , en se servant du numéro d'identification unique du déclarant. Le déclarant sera inclus dans l'échantillon de première phase si $R_i \leq (0, v_g^*)$. Pour ce qui est du tirage de l'échantillon de la seconde phase, on se sert aussi d'un échantillonnage de Bernoulli, mais avec un ensemble de nombres pseudo-aléatoires différents. L'échantillonnage de Bernoulli permet de procéder à l'échantillonnage et au traitement avant d'avoir toute l'information voulue sur l'univers du déclarant. Cet avantage est précieux car la collecte d'informations s'échelonne normalement sur deux ans. Avec l'échantillonnage de Bernoulli, la taille des échantillons est aléatoire. Dans Choudhry, Lavallée et Hidiroglou (1989b), on calcule la variance de $X_{h-STRAT} = \sum_g \sum_{i \in s \cap h} Y_i / (v_g^* \cdot v_{gh}^*)$ au moyen d'un échantillon aléatoire simple qui tient lieu d'échantillon de Bernoulli (voir Sunter 1986). Suivant cette méthode, on prélève à la première phase un échantillon aléatoire simple de taille fixe $n_g^* = v_g^* \cdot N_g^*$ dans la strate g . Désignons par n_{gh}^* le nombre d'unités de la strate g (taille) dans l'échantillon de première phase qui font partie de la strate h (CTI4). Dans la seconde phase, on prélève un échantillon aléatoire simple de taille $n_{gh}^* = v_{gh}^* \cdot n_{gh}^*$ dans la strate h et la strate g , v_{gh}^* étant considérée fixe. La variance de $X_{h-STRAT}$ est définie

$$V_h = \sum^g \left(\frac{1}{v_g^* \cdot v_{gh}^*} - 1 \right) \cdot A_{gh} + \sum^g \left(\frac{1}{v_g^*} - 1 \right) \cdot B_{gh},$$

où

$$A_{gh} = N_{gh}^* \cdot S_{gh}^2,$$

$$B_{gh} = \left(\frac{N_g^*}{N_{gh}^*} - 1 \right) \cdot \left(\frac{Y_{gh}^2}{N_{gh}^*} - S_{gh}^2 \right),$$

et S_{gh}^2 est la variance dans la strate de seconde phase gh (taille \times CTI4).

Voici comment est structurée cet article. Dans la section 2, nous exposons le problème de la répartition optimale par rapport à l'échantillonnage à deux phases de dossiers fiscaux. Nous proposons une solution itérative appelée 'méthode exacte'. Dans la section 3, nous décrivons une version approchée de la répartition optimale qui peut servir à générer des valeurs de départ pour la méthode exacte. La section 4 contient les résultats d'une étude empirique où l'on compare diverses valeurs de départ pour la méthode exacte. Enfin, la section 5 sert de conclusion.

Une méthode de répartition de l'échantillon pour des plans d'échantillonnage à deux phases

J.B. ARMSTRONG et C.F.J. WU¹

RÉSUMÉ

Dans la foulée de l'élaboration d'un plan de sondage pour des enquêtes-entreprises à Statistique Canada, nous exposons le problème de la répartition de l'échantillon pour un plan de sondage général à deux phases comme un problème de programmation non linéaire sous contrainte. En exploitant la structure mathématique du problème, nous proposons une solution qui consiste en des itérations entre deux sous-problèmes qui sont beaucoup moins complexes sur le plan du calcul. Lorsqu'on utilise une solution approximative comme valeur de départ, la méthode proposée donne des résultats très satisfaisants dans une étude empirique.

MOTS CLÉS: Répartition optimale; programmation convexe.

1. INTRODUCTION

Le but de cet article est de proposer une méthode de répartition de l'échantillon pour des plans de sondage à deux phases. Supposons qu'il faille répartir en L strates une population de taille N en fonction d'une variable auxiliaire z dont nous ne connaissons pas la valeur avant l'échantillonnage. En revanche, nous connaissons la valeur d'une deuxième variable auxiliaire (taille), x , qui est corrélée avec la variable étudiée, y , pour toutes les unités de la population. Dans la première phase de l'échantillonnage, la population est divisée en G strates selon x . On tire tout d'abord un échantillon aléatoire simple de la strate g ($g = 1, 2, \dots, G$) avec comme fraction de sondage v_g , et on observe la valeur z pour chaque unité échantillonnée. Dans la seconde phase, on tire un nouvel échantillon parmi les unités qui forment l'échantillon de la strate g et pour lesquelles la valeur z appartient à la classe h ($h = 1, 2, \dots, L$); la fraction de sondage est alors v_{gh} . On observe ensuite la valeur y pour les unités de l'échantillon de la

seconde phase. Dans le cas où il n'y a pas de stratification en fonction de la taille (c.-à-d. $G = 1$), Cochran (1977) définit la formule de répartition qui réduit au minimum la variance de l'estimation $Y = \sum_h \sum_{i \in s \cap h} y_i / (v \cdot v_h)$ du total de population $Y = \sum_h N_h \cdot \bar{y}_h$, étant donné un coût fixe, C , pour l'enquête, N_h et \bar{y}_h étant, respectivement, la taille et la moyenne de la population pour la strate h et $\sum_{i \in s \cap h} y_i$ désignant la somme des valeurs de y pour les unités de l'échantillon de la seconde phase, s_2 , pour lesquelles la valeur z est incluse dans la classe h . Si les estimations d'enquête servent à l'analyse, la variance du total estimé pour la classe h , $\bar{Y}_h = \sum_{i \in s \cap h} y_i / (v \cdot v_h)$, mérite aussi d'être considérée. Sedransk (1965), Booth et Sedransk (1969), Rao (1973) et Smith (1989) se penchent sur des problèmes de répartition où l'on minimise une fonction de variances de totaux de classe estimés, étant donné une contrainte de coût. La méthode que nous décrivons dans cet article peut servir à résoudre le problème de la répartition dans tous les cas de stratification où la variance du total estimé pour chaque classe z est assujettie à une contrainte. L'idée de cette méthode nous est venue à la suite d'une expérimentation faite au cours d'une enquête-entreprises de Statistique Canada, où étaient échantillonnées des dossiers fiscaux d'entreprises.

¹ J.B. Armstrong, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6 et C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (Ontario) N2L 3G1.

7. CONCLUSIONS

Les travaux continus sur les techniques d'estimation de la variance dont il est fait état dans le présent article visent à étendre l'analyse: 1) aux méthodes d'imputation fondées sur un modèle qui sont seulement implicites, en particulier la méthode d'imputation du plus proche voisin; 2) au cas où plusieurs méthodes d'imputation sont utilisées dans la même étude.

Deville et Särndal (1992) présentent les résultats d'une extension dans laquelle l'estimateur de Horwitz-Thompson $\hat{t} = \sum^s y_k/\pi_k$ sert de prototype. L'estimateur utilisant les données après imputation est alors donné par

$$\hat{t} = \sum^r y_k/\pi_k + \left(\sum^{s-r} x_k/\pi_k \right) \hat{b} = \sum^s y_k/\pi_k - \sum^{s-r} e_k/\pi_k,$$

où $e_k = y_k - x_k$ est le résidu d'imputation pour l'unité k , obtenu par régression multiple.

REMERCIEMENTS

J'exprime ma reconnaissance à M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee et G. Reinhard de Statistique Canada pour leur collaboration aux travaux ayant mené à cet article. Les commentaires de deux arbitres ont permis d'améliorer le manuscrit initial, et ont été vivement appréciés.

BIBLIOGRAPHIE

DEVILLE, J. C., et SÄRNDAL, C.-E. (1992). Variance estimation for survey data with regression imputation. Rapport technique.

HERZOG, T. N., et RUBIN, D. B. (1983). Using multiple imputations to handle nonresponse in surveys. Dans *Incomplete Data in Sample Surveys*, (Eds. W. G. Madow, I. Olkin et D. B. Rubin). New York: Academic Press, 209-245.

FAY, R. E. (1991). A design-based perspective on missing data variance. Proceedings, 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.

LEE, H., RANCOURT, E., et SÄRNDAL, C.-E. (1992). Experiments with variance estimation from survey data with imputed values. Rapport, Division des méthodes d'enquêtes-entreprises, Statistique Canada, soumis pour publication.

LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys (avec discussion). *Journal of Business and Economic Statistics*, 6, 287-301.

PRITZKER, L., OGUS, J., et HANSEN, M. H. (1965). Computer editing methods: some applications and results. *Bulletin de l'Institut International de Statistique*, 41, 442-466.

RAO, J. N. K. (1990). Variance estimation under imputation for missing data. Ce manuscrit a été lu avec la gracieuse permission de l'auteur.

RAO, J. N. K. (1992). Jackknife variance estimation under imputation for missing survey data. Ce manuscrit a été lu avec la gracieuse permission de l'auteur.

RUBIN, D. B. (1986). Initiation à l'imputation multiple pour les cas de non-réponse. *Techniques d'enquête*, 12, 41-52.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SÄRNDAL, C.-E. (1990). Estimation of precision in the generalized estimation system when imputation is used. Rapport, Secteur de l'informatique et de la méthodologie, Statistique Canada, mars 31, 1990.

Le biais relatif d'un estimateur V a été calculé sous la forme $\{\text{moy}(V) - \text{var}(\hat{t}_\bullet)\} / \text{var}(\hat{t}_\bullet)$, où moy(V) est la moyenne des 100,000 valeurs de V , et (\hat{t}_\bullet) est la variance des 100,000 valeurs de \hat{t}_\bullet . La simulation montre que l'estimateur de la variance fondé sur un modèle $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$ est quasi non biaisé pour les trois mécanismes de réponse. D'une certaine façon, ce n'est pas surprenant parce que la population a été créée de manière à concorder avec le modèle d'imputation par le quotient. Les mécanismes 1 et 2 sont du type non neutre et ne répondent pas à la condition (a) de la section 4, requise pour que V_{tot} soit dépourvu de biais. Fait intéressant, toutefois, le biais de V_{tot} dans cet exemple demeure quand même faible. L'estimateur à deux phases fonctionne bien pour le mécanisme 3 (réponse uniforme), cas pour lequel il a été conçu; dans les autres cas, il est biaisé. Enfin, considérer les valeurs imputées comme des données réelles entraîne, comme prévu, une sous-estimation prononcée de la vraie variance pour les trois mécanismes. Une étude de Monte Carlo plus approfondie de l'estimation par le quotient est décrite dans Lee, Rancourt et Särndal (1992). Cet article donne une idée de l'effet d'une définition déficiente du modèle d'imputation, sujet également abordé dans Rao (1992).

6. VALEURS IMPUTÉES AYANT UN RÉSIDU AJOUTÉ

Nous pouvons distinguer deux types de valeurs imputées: 1) la valeur imputée $y_{\text{imp},k}$ est constituée d'une valeur prévue seulement, $y_{\text{pred},k}$, par exemple quand il s'agit d'une valeur provenant d'une ligne ou d'une surface résultant d'un ajustement de régression. Par exemple, dans la méthode d'imputation par le quotient utilisée ci-dessus, $y_{\text{imp},k} = y_{\text{pred},k} = Bx_k$, où $B = (\sum y_k x_k) / (\sum x_k)$; 2) la valeur imputée $y_{\text{imp},k}$ est constituée d'une valeur prévue et d'un résidu, c.-à-d. que $y_{\text{imp},k} = y_{\text{pred},k} + e_k^*$. Le terme résiduel, dont le rôle est de rendre les valeurs imputées plus conformes aux observations réelles, peut être obtenu par échantillonnage des résidus $e_k = y_k - y_{\text{pred},k}$ calculés pour les unités répondantes $k \in r$. Un moyen de le faire est exposé ci-dessous. Ce type d'imputation est parfois recommandée, dans la littérature, comme moyen de préserver les distributions des valeurs imputées; voir, par exemple, l'analyse faite dans Little (1988). Le processus d'imputation devient alors plus ardu à réaliser et, pour les besoins du SGE (dont le but principal est une estimation valide de la précision des estimations de l'enquête), il n'est pas clair que les avantages réalisés justifient l'effort additionnel.

Nous présentons cependant une marche à suivre pour l'imputation par "valeur prévue plus résidu" dans le cas où le modèle d'imputation par le quotient courant est pris comme point de départ: Pour $k \in r$, calculons $e_k = y_k - Bx_k$, où $B = (\sum y_k x_k) / (\sum x_k)$, puis $e_k = e_k / \sqrt{x_k}$. On obtient ainsi un ensemble de m "résidus standardisés" e_k . Ensuite, pour une unité $k \in s - r$, calculons $e_k^* = \sqrt{x_k} e_k$, où e_k^* est tiré par EASAR de l'ensemble, et x_k appartient à l'unité exigeant une imputation. Les unités ayant une valeur x élevée ont alors tendance à obtenir des résidus e_k^* plus élevés, ce qui est conforme au modèle. Posons ensuite $e_k^* = e_k^* - (\sum_{s-r} e_k^*) / (n - m)$. Pour $k \in s - r$, imputons $y_{\text{imp},k} = Bx_k + e_k^*$, $k \in s - r$; pour $k \in r$, nous avons les observations réelles, y_k . Puisque la somme des e_k^* doit par définition évaluer zéro pour l'ensemble des unités de $s - r$, l'estimateur ponctuel est donné par $\hat{t}_\bullet = (N/n) \sum_{s-r} y_{\text{imp},k} = N \bar{x}_s \bar{y}_r / x_r$, comme à la section 5, mais sa variance est différente. On peut montrer que $E_{\hat{t}_\bullet} E_{\hat{t}_\bullet} (S_{y_{\text{imp},k}}^2 - S_{y_s}^2) \approx 0$, où $E_{\hat{t}_\bullet}$ désigne la moyenne par rapport à la sélection au hasard d'un résidu standardisé. Autrement dit, la différence entre la variance calculée à partir des données après imputation, $S_{y_{\text{imp},k}}^2$, et la variance inconnue d'un échantillon formé entièrement d'observations réelles, $S_{y_s}^2$, est approximativement égale à zéro en moyenne. Nous pouvons utiliser $V_{\text{sam}} = N^2 (1/n - 1/N) S_{y_{\text{imp},k}}^2$ en tant qu'estimateur à peu près globalement non biaisé de la composante variance d'échantillonnage. Il n'est pas nécessaire maintenant d'ajouter une correction V_{diff} . Toutefois, il reste toujours à calculer un estimateur de la variance d'imputation $V_{\text{imp}} = N^2 (1/m - 1/n) C_1 \sigma^2$ et à l'ajouter à V_{sam} .

Si m n'est pas trop petit, les approximations $\hat{\sigma}^2 \approx (\sum r_k e_k^2)/(\sum r_k x_k)$ où $e_k = y_k - Bx_k$ et $C_0 \approx (1 - m/n)^{x_s - r}$ sont suffisamment bonnes pour la plupart des applications. La composante variance d'imputation peut être ainsi exprimée

$$V_{\text{imp}} = N^2(1/m - 1/n)Ax_s^2,$$

où $A = x_s - r/x_r$. La constante A reflète l'effet de sélection dû à la non-réponse. Si les unités ayant une valeur élevée sont moins enclines à répondre que les unités ayant une valeur faible, A peut être beaucoup plus grand que un et, pour un échantillon donné s et un nombre de répondants donné m , la composante V_{imp} tend à être élevée, par rapport au cas où, disons, toutes les unités sont également enclines à répondre. Cette tendance apparaît raisonnable intuitivement. Deux cas spéciaux sont à signaler: 1) Si tous les x_k sont égaux à 1, l'estimation de la variance totale devient simplement

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}} = N^2(1/m - 1/n)S_y^2,$$

où S_y^2 est la variance des m observations réelles y_k . Cela correspond à la variance obtenue selon un plan d'échantillonnage à deux phases, chaque phase étant fondée sur un EASSR. 2) Si aucune imputation n'est nécessaire, c.-à-d. si $s = r$, on a $V_{\text{imp}} = 0$, et

$$V_{\text{tot}} = V_{\text{sam}} = N^2(1/n - 1/N)S_y^2.$$

Autrement dit, notre méthode donne l'estimateur bien connu de la variance pour un EASSR.

Une étude de Monte Carlo comportant 100,000 ensembles de réponses répétés r a été effectuée afin de confirmer les résultats ci-dessus pour une imputation par le quotient courant. Une population finie de taille $N = 100$ a été créée conformément au modèle décrit par (5.1) et (5.2). L'ensemble de réponses typique r a été obtenu de la façon suivante: tirage par EASSR d'un échantillon s de taille $n = 30$, s étant donné, création de r par un mécanisme de réponse prenant la forme d'essais de Bernoulli indépendants, un pour chaque k es, avec une probabilité θ_k que le résultat soit "réponse". Trois mécanismes de réponse différents ont été utilisés; mécanisme 1: θ_k s'accroît avec y_k de telle sorte que $\theta_k = 1 - \exp(-a_1 y_k)$; mécanisme 2: θ_k s'accroît à mesure que y_k décroît, de telle sorte que $\theta_k = \exp(-a_2 y_k)$; mécanisme 3: θ_k est constant à 0.7, c.-à-d. que le mécanisme de réponse est uniforme. Les constantes a_1 et a_2 dans les deux premiers mécanismes de réponse (qui peuvent être qualifiées de non neutres) ont été fixées de manière à produire une probabilité de réponse moyenne de 0.7. Les tailles des ensembles de réponses r réalisées ont donc varié autour d'une moyenne de 21 pour les trois mécanismes. Pour chaque r , l'estimation ponctuelle \hat{t} , donnée par (5.3) a été calculée, ainsi que trois estimateurs de la variance différents, $V = V(\hat{t})$. Ces estimateurs étaient: 1) l'estimateur de la variance fondé sur un modèle $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$, égal au total de (5.4) et de (5.5); 2) l'estimateur de la variance pour un échantillonnage à deux phases $N^2(1/n - 1/N)S_y^2 + N^2(1/m - 1/n)$ la variance pour un échantillonnage à deux phases $N^2(1/n - 1/N)S_y^2 + N^2(1/m - 1/n)$ initial (Rao 1990); et 3) l'estimateur de la variance non corrigé ordinaire $N^2(1/n - 1/N)S_y^2$, obtenu en faisant comme si les valeurs imputées étaient aussi bonnes que les valeurs réelles. Les résultats figurent dans le tableau qui suit:

Estimateur de V	Biais relatif de V en %		
	Mécanisme 1	Mécanisme 2	Mécanisme 3
Fondé sur un modèle	-0.20	-4.64	-3.99
À deux phases	9.95	-12.49	-1.11
Non corrigé ordinaire	-25.73	-37.90	-33.21

L'imputation par le quotient courant se fonde sur le modèle

$$(5.1)$$

$$y_k = \beta x_k + \epsilon_k,$$

où les ϵ_k sont des erreurs du modèle non corrélées telles que

$$(5.2)$$

$$E_{\xi}(\epsilon_k) = 0, \quad V_{\xi}(\epsilon_k) = \sigma^2 x_k.$$

Supposons que l'échantillon s soit choisi par EASSR. Désignons les tailles respectives de s , r et $s - r$ par n , m et $n - m$. Si aucune imputation n'était nécessaire, l'estimateur de $t = \sum_U y_k$ serait $\hat{t} = N y_s$. En utilisant les données après imputation, nous obtenons

$$(5.3)$$

$$\hat{t} \bullet = (N/n) \sum_s y_{\bullet k} = N \bar{x}_s \bar{y}_r / \bar{x}_r.$$

(Le trait supérieur et l'indice s , r ou $s - r$ désignent la "moyenne ordinaire", par exemple $\bar{y}_r = \sum r y_k / m$, $\bar{x}_{s-r} = \sum_{s-r} x_k / (n - m)$, etc.) D'après les résultats de la section précédente, nous avons $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$, avec $V_{\text{sam}} = E_{\xi} \{ N^2 (1/n - 1/N) S_y^2 \}$ et $V_{\text{imp}} = E_s E_r \{ N^2 (1/m - 1/n) C_1 \sigma^2 \}$, où $S_y^2 = \sum_U (y_k - \bar{y}^U)^2 / (N - 1)$ et $C_1 = \bar{x}_s \bar{x}_{s-r} / \bar{x}_r$, une constante connue. Le terme mixte (4.2) égale exactement zéro dans ce cas. Notre méthode d'estimation de la variance donne $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$ avec

$$(5.4)$$

$$V_{\text{sam}} = N^2 (1/n - 1/N) \{ S_{y_{\bullet s}}^2 + C_0 \hat{\sigma}^2 \},$$

$$(5.5)$$

$$V_{\text{imp}} = N^2 (1/m - 1/n) C_1 \hat{\sigma}^2,$$

où $S_{y_{\bullet s}}^2 = \sum_s (y_{\bullet k} - \bar{y}_{\bullet s})^2 / (n - 1)$ est la variance calculée à partir des données après imputation, et nous avons choisi d'estimer σ^2 par la formule non biaisée selon le modèle

$$\hat{\sigma}^2 = \frac{1}{\sum_r (y_k - \bar{b} x_k)^2} = \frac{x_r \{ 1 - (1/m) (cv_{x_r})^2 \}}{m - 1},$$

où $cv_{x_r} = S_{x_r} / \bar{x}_r$ est le coefficient de variation de x dans l'ensemble de réponses r . La constante C_0 est donnée par

$$C_0 = \frac{1}{\hat{\sigma}^2} E_{\xi} (S_y^2 - S_{y_{\bullet s}}^2),$$

où

$$S_{y_s}^2 = \frac{1}{n - 1} \sum_s (y_k - \bar{y}_s)^2$$

est la variance de l'échantillon (inconnue) selon des données formées de 100 % d'observations réelles. Après évaluation, on a

$$C_0 = \frac{1}{n - 1} \left\{ \sum_{s-r} x_k - \frac{\sum_r x_k^2}{\sum_{s-r} x_k^2} + \frac{1}{n} \frac{\sum_r x_k}{\sum_{s-r} x_k} \frac{\sum_s x_k}{\sum_{s-r} x_k} \right\}.$$

un profil de réponse systématique en vertu duquel les unités ayant des valeurs x_k élevées sont moins susceptibles de répondre que les unités ayant des valeurs x_k faibles. Si les probabilités de réponse dépendent explicitement des valeurs y_k , la situation est différente; le mécanisme de réponse n'est pas neutre et la condition (a) n'est pas remplie. Il y aura alors un biais dans V_{tot} attribuable à cette non-neutralité; les simulations décrites dans Lee, Rancourt et Särndal (1992) donnent une idée de l'ampleur de ce biais.

Exemple. On utilise l'EASSR pour prélever l'échantillon s , formé de n unités parmi N . Désignons par m la taille de l'ensemble de réponses r . Supposons qu'on impute la moyenne des répondants aux unités exigeant une imputation. Le modèle d'imputation correspondant à $\hat{y}_k(\epsilon_k) = \sigma^2$. Autrement dit, $y_{\bullet k} = y_k$ si $k \in r$ et $y_{\bullet k} = \beta = \bar{y}_r$ si $k \in s - r$, et nous obtenons $\hat{y}_{\bullet k} = (N/n) \sum_s y_{\bullet k} = N \bar{y}_r$. Ici, l'estimateur ordinaire, fondé sur le plan, de la variance pour une réponse de 100 % est $V_p^2 = N^2(1/n - 1/N) \sum_s (y_k - \bar{y}_s)^2 / (n - 1)$; lorsque cette formule est appliquée aux données après imputation, on obtient $V_{\bullet p}^2 = N^2(1/n - 1/N) \{ (m - 1) / (n - 1) \} S_{y_r}^2$, où $S_{y_r}^2 = \sum_r (y_k - \bar{y}_r)^2 / (m - 1)$. D'autres calculs donnent $V_{\text{dif}}^2 = N^2(1/n - 1/N) \{ (n - m) / (n - 1) \} S_{y_r}^2$ et $V_{\text{imp}}^2 = N^2(1/m - 1/n) S_{y_r}^2$. Ainsi, $V_{\text{sam}}^2 = V_{\bullet p}^2 + V_{\text{dif}}^2 = N^2(1/n - 1/N) S_{y_r}^2$ et $V_{\text{tot}}^2 = N^2(1/m - 1/N) S_{y_r}^2$, lequel est facile à accepter comme "bon" estimateur de la variance pour cette règle d'imputation simple. Le tableau qui suit montre la contribution de chacun des trois termes à l'estimateur de la variance totale V_{tot} , pour différents taux d'imputation, en supposant que N est grand comparativement à m et à n , et que $(m - 1) / m \approx (n - 1) / n \approx 1$.

taux d'imputation en %	contribution en % à V_{tot}		
	$V_{\bullet p}^2$	V_{dif}^2	V_{imp}^2
100 (1 - m/n)	10	81	9
30	10	64	21
20	10	49	30

Le tableau illustre les dangers de considérer les valeurs imputées comme des données réelles. Avec 30 % de valeurs imputées, l'estimateur de la variance ordinaire $V_{\bullet p}^2$ couvre, dans cet exemple, moins de la moitié de la variance totale correctement estimée. L'imputation par la moyenne des répondants est un exemple utile, du fait de la relative simplicité des résultats. Mais habituellement, en situation réelle, une telle imputation n'est ni justifiée ni efficace. Le modèle sous-jacent n'est pas assez raffiné pour éviter une erreur systématique dans les estimations ponctuelles, et les résidus $e_k = y_k - \bar{y}_r$ peuvent varier considérablement.

5. APPLICATION À L'IMPUTATION PAR LE QUOTIENT COURANT

La méthode se fonde sur l'hypothèse qu'une valeur auxiliaire positive x_k est connue pour chaque unité $k \in s$. Si $k \in s - r$, nous imputons $y_{\text{imp},k} = Bx_k$ où $B = (\sum_r y_k) / (\sum_r x_k)$. Les données après imputation sont les suivantes:

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ Bx_k & \text{si } k \in s - r. \end{cases}$$

$$E_{\xi}(V_p - V_{\diamond p} | s, r) = V_{\text{diff}}.$$

Alors, pour s et r donnés, trouvons un estimateur non biaisé selon le modèle, désigné V_{diff}^{ξ} . Cela exigera normalement l'estimation de certains paramètres du modèle ξ . Par conséquent,

$$E_{\xi}(V_{\text{diff}}^{\xi} | s, r) = E_{\xi}(V_p - V_{\diamond p} | s, r).$$

Alors

$$V_{\text{sam}}^{\xi} = V_{\diamond p}^{\xi} + V_{\text{diff}}^{\xi}$$

est globalement non biaisé pour la composante $V_{\text{sam}}^{\xi} = E_{\xi}V_p$, comme le montre le calcul qui suit:

$$E_{\xi}E_sE_r(V_{\text{sam}}^{\xi}) = E_sE_r\{E_{\xi}(V_{\diamond p}^{\xi}) + E_{\xi}(V_{\text{diff}}^{\xi})\}$$

$$= E_sE_r\{E_{\xi}(V_p^{\xi})\} = E_{\xi}E_s(V_p^{\xi})$$

$$= E_{\xi}V_p^{\xi} = V_{\text{sam}}^{\xi}.$$

(iii) Estimation de la composante variance d'imputation. Il suffit de trouver un estimateur, $V_{\xi c}^{\xi}$, qui soit non biaisé selon le modèle pour $V_{\xi c}^{\xi}$. Autrement dit, $E_{\xi}(V_{\xi c}^{\xi}) = V_{\xi c}^{\xi}$. Encore ici, l'estimation de paramètres inconnus du modèle ξ peut être nécessaire. Il en résulte que $V_{\xi c}^{\xi}$ est globalement non biaisé pour la composante variance d'imputation V_{imp}^{ξ} , puisque

$$E_sE_rE_{\xi}(V_{\xi c}^{\xi}) = E_sE_rV_{\xi c}^{\xi} = V_{\text{imp}}^{\xi}.$$

Enfin, un estimateur globalement non biaisé de V_{tot}^{ξ} est donné par

$$V_{\text{tot}}^{\xi} = V_{\text{sam}}^{\xi} + V_{\text{imp}}^{\xi},$$

où $V_{\text{sam}}^{\xi} = V_{\diamond p}^{\xi} + V_{\text{diff}}^{\xi}$ et $V_{\text{imp}}^{\xi} = V_{\xi c}^{\xi}$. Notons que le rôle de V_{diff}^{ξ} est d'apporter une correction tenant compte du fait que les données après imputation peuvent afficher une variation "moindre que naturelle". Cela se produit souvent lorsque $y_{\text{imp},k}^{\xi}$ égale la valeur prévue provenant d'un ajustement de régression, c.-à-d. "la valeur située sur la ligne". La variation autour de la ligne n'est pas reflétée par la valeur prévue.

Pour que l'estimateur V_{tot}^{ξ} construit ci-dessus soit globalement non biaisé, la condition (a) doit être remplie, le terme énoncé en (4.2) doit être égal à zéro, et le modèle d'imputation doit être approprié, de façon que V_{diff}^{ξ} et $V_{\xi c}^{\xi}$ soient non biaisés selon le modèle pour V_{diff}^{ξ} et $V_{\xi c}^{\xi}$, respectivement. Il se peut que de légers écarts par rapport au modèle d'imputation supposé n'aient pas de conséquences graves, mais si la définition du modèle d'imputation est grossièrement déficiente, il est clair que V_{tot}^{ξ} pourra être gravement biaisé, en raison du biais selon le modèle de V_{diff}^{ξ} et $V_{\xi c}^{\xi}$. Des simulations de Monte Carlo décrites dans Lee, Rancourt et Särndal (1992) montrent que l'estimateur de la variance V_{tot}^{ξ} est passablement robuste à l'égard des déficiences du modèle d'imputation. Quoi qu'il en soit, le fait d'ajouter les termes V_{diff}^{ξ} et $V_{\xi c}^{\xi}$ constitue une importante amélioration par rapport à l'utilisation de l'estimateur de variance non corrigé élémentaire $V_{\diamond p}^{\xi}$.

Notons que si le modèle d'imputation demeure valable, une estimation non biaisée de la variance est fournie par la méthode même si les probabilités de réponse diffèrent entre les unités, pourvu qu'elles dépendent des valeurs x_k seulement. Autrement dit, nous pouvons appliquer

Examinons maintenant la variance globale donnée par

$$V_{\text{tot}} = E_{\xi} E_s E_r \{ (t_{\bullet} - t)^2 \},$$

qui peut être aussi appelée variance prévisible en vertu du modèle d'imputation ξ . Nous obtenons:

$$\begin{aligned} V_{\text{tot}} &= E_{\xi sr} \{ (t_{\bullet} - t)^2 \} \\ &= E_{\xi} E_s E_r \{ (t - t) + (t_{\bullet} - t) \}^2 \\ &= E_{\xi} V_p + E_s E_r V_{\xi c}, \end{aligned} \quad (4.1)$$

où $V_p = E_s \{ (t - t)^2 \}$ est la variance fondée sur le plan de t , en supposant que t est non biaisé selon le plan pour le total t . (Pour un estimateur légèrement biaisé selon le plan, V_p est l'erreur quadratique moyenne selon le plan de t .) Notons que $(t - t)$ dépend de s seulement, et non de r . De plus,

$$V_{\xi c} = E_{\xi} \{ (t_{\bullet} - t)^2 \mid s, r \}$$

est la variance (selon le modèle) de l'erreur d'imputation, étant donné des ensembles s et r . L'indice c signifie "conditionnelle". Pour en arriver à (4.1), on suppose que la condition (a) est vérifiée, c.-à-d. que l'espérance E_{ξ} peut être déplacée à l'intérieur de $E_s E_r$, et que le terme mixte

$$2E_{\xi} E_s [(t - t) \{ E_r(t_{\bullet} - t) \mid s \}] \quad (4.2)$$

disparaît, ou est suffisamment proche de zéro pour être omis. Ce serait le cas si l'erreur d'imputation probable était nulle ou négligeable en vertu du mécanisme de réponse, étant donné l'échantillon probabiliste s réalisé. Même si (4.2) n'est pas exactement zéro pour le mécanisme déterminant la réponse, on peut souvent assimiler (4.2) à zéro et utiliser quand même la méthode ci-dessous, laquelle permet d'obtenir une bien meilleure estimation de la variance que si l'on prétendait naïvement que les données imputées sont aussi valables que les données réellement observées. Pour une imputation par le quotient et un EASSR, cas que nous examinons à la section 5, le terme (4.2) égale exactement zéro.

En utilisant les désignations $V_{\text{sam}} = E_{\xi} V_p$ et $V_{\text{imp}} = E_s E_r V_{\xi c}$ dans (4.1), on obtient

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$$

ou

$$\text{variance totale} = \text{variance d'échantillonnage} + \text{variance d'imputation}.$$

L'objectif est d'estimer la variance totale, de façon à déterminer un intervalle de confiance valide pour la valeur inconnue t . Notre méthode consiste à obtenir des estimations distinctes, V_{sam} et V_{imp} , des deux composantes $V_{\text{sam}} = E_{\xi} V_p$ et $V_{\text{imp}} = E_s E_r V_{\xi c}$. Les données disponibles pour cette estimation sont les $y_{\bullet k}$, $k \in s$. La façon d'obtenir V_{sam} et V_{imp} est la suivante:

(i) Estimation de la composante d'échantillonnage. Soit V_p l'estimateur ordinaire (non biaisé selon le plan ou quasi non biaisé selon le plan) de la variance selon le plan V_s . Appelons V_p la quantité résultant du calcul de V_p à partir des valeurs après imputation $y_{\bullet k}$, $k \in s$. Pour de nombreuses règles d'imputation, $V_{\bullet p}$ sous-estime V_{sam} . La sous-estimation est compensée de la façon suivante. Evaluons l'espérance conditionnelle

2) Ajouter un terme de correction tenant compte du fait que de nombreuses règles d'imputation produisent des données ayant une variabilité "moindre que naturelle", phénomène propre à entraîner une sous-estimation de la variance d'échantillonnage à moins que des mesures correctives soient prises. Enfin, la composante R_{imp}^{sam} est facilement calculée à partir des données après imputation. L'utilisateur acceptera facilement l'argument selon lequel la variance fournie par la formule ordinaire n'est pas suffisante à elle seule; quelque chose doit être ajoutée pour tenir compte de l'imperfection de la règle d'imputation.

Une propriété intéressante de la méthode est que si aucune imputation n'est nécessaire, c.-à-d. si $r = s$, il en résulte que $R_{imp}^{sam} = 0$ et que R_{sam}^{sam} égale l'estimateur ordinaire de la variance" qui aurait été utilisé si 100 % des valeurs avaient été réellement observées.

4. EXPOSE THEORIQUE

L'erreur totale de t_{\bullet} est ainsi décomposée:

$$t_{\bullet} - t = (t - t) + (t_{\bullet} - t) = \text{erreur d'échantillonnage} + \text{erreur d'imputation}.$$

L'erreur d'imputation est la différence entre l'estimation inconnue qui aurait été calculée si les données avaient été formées entièrement d'observations réelles, et l'estimation qui peut être calculée à partir des données après imputation. L'erreur d'imputation est

$$t_{\bullet} - t = - \sum_{s-r} w_k e_k, \quad \text{ou}$$

$$e_k = y_k - y_{imp,k}$$

est un **résidu d'imputation** qui ne peut être observé pour une unité $k \in s - r$. La grandeur de e_k dépend de la qualité d'ajustement offerte par le modèle d'imputation. Les résidus sont faibles si la méthode d'imputation donne des valeurs substitués presque parfaites. L'analyse, à partir d'ici, peut prendre diverses directions. Nous utiliserons une méthode **fondée sur un modèle** dans laquelle trois différentes distributions de probabilité sont examinées. Les symboles d'espérance correspondants sont E_{ξ} , E_s et E_r . L'indice ξ signifie "par rapport au modèle d'imputation", l'indice s signifie "par rapport au plan d'échantillonnage" et l'indice r signifie "par rapport au mécanisme de réponse, s étant donné". Le modèle découle de la règle d'imputation, et est donc connu; le plan d'échantillonnage est la distribution d'échantillonnage probabiliste donnée, et est donc aussi connu; le mécanisme de réponse est une distribution habituellement inconnue régissant la réponse, l'échantillon s étant donné.

L'estimateur t_{\bullet} est globalement non biaisé, au sens où $E_{\xi} E_s E_r (t_{\bullet} - t) = 0$, si deux conditions sont remplies:

a) l'ordre des opérateurs d'espérance peut être modifié, de sorte que $E_{\xi} E_s E_r (\cdot)$ peut être évalué sous la forme $E_s E_r \{ E_{\xi} (\cdot | s, r) \}$, et

b) le résidu d'imputation $e_k = y_k - y_{imp,k}$ a une espérance (selon le modèle) nulle pour chaque $k \in r$, c.-à-d. $E_{\xi}(e_k) = 0$, d'où il résulte que $E_{\xi}(t_{\bullet} - t) = 0$.

La condition (a) est remplie si la réponse est régie par un mécanisme qui peut dépendre de s et de données auxiliaires, mais non des y_k , $k \in s$. Autrement dit, la probabilité $q(r)$ de réaliser l'ensemble de réponses r est de la forme $q(r) = q(r | s, \{x_k: k \in s\})$, où $\{x_k: k \in s\}$ désigne les données auxiliaires. Le mécanisme de réponse peut alors être qualifié de neutre (c.-à-d. qu'on peut ne pas en tenir compte).

Soit $U = \{1, \dots, k, \dots, N\}$ une population finie; supposons que y désigne une des variables étudiées dans l'enquête. L'objectif est d'estimer le total de y pour l'ensemble de la population, c.-à-d. $t = \sum U y_k$. (Pour n importe quel ensemble C d'unités de la population, $C \subseteq U$, $\sum_C y_k$ est une désignation abrégée de $t = \sum U y_k$; par exemple, signifie $\sum_{k \in U} y_k$.) Un échantillon probabiliste s est prélevé selon un plan d'échantillonnage déterminé. Les probabilités d'inclusion sont connues, et l'on obtiendrait des estimations ordinaires, fondées sur le plan, de la variance si toutes les unités $k \in s$ étaient observées. Toutefois, il y a des données manquantes. Soit r le sous-ensemble de s pour lequel les valeurs y_k sont réellement observées. Pour le complément, $s - r$, des valeurs imputées sont calculées. Les **données après imputation** sont l'ensemble des valeurs dénotées $y_{\star k}$, $k \in s$, telles que:

$$y_{\star k} = \begin{cases} y_k & \text{si } k \in r \\ y_{\text{imp},k} & \text{si } k \in s - r, \end{cases}$$

où y_k est une valeur réellement observée, et $y_{\text{imp},k}$ est la valeur imputée pour l'unité k . Si $r = s$, il n'y a aucune imputation; toutes les données sont des observations réelles. Écrivons l'estimateur de t qui serait utilisé dans le cas d'une réponse de 100% (c.-à-d. $r = s$) de la façon suivante: $t = \sum_{k \in s} w_k y_k$, où w_k est le poids attribué à l'observation y_k . Par exemple, pour un échantillonnage aléatoire simple sans remise (EASSR) de n unités parmi N , $w_k = N/n$ pour tous les $k \in s$ lorsque la moyenne de l'échantillon étendue à la population est utilisée pour estimer t , et $w_k = (\bar{z}_U / \bar{z}_s) (N/n) = (\sum U z_k) / (\sum s z_k)$ pour tous les $k \in s$ lorsqu'on utilise l'estimateur par quotient avec z comme variable auxiliaire.

Lorsque les données contiennent des valeurs imputées, l'estimateur de t est $t_{\star} = \sum s w_k y_{\star k}$. Autrement dit, nous supposons que les poids w_k sont identiques à ceux qui sont utilisés lorsque toutes les données sont des observations réelles. Ce principe est utilisé dans les modules d'estimation du SGE. Il s'appuie sur l'hypothèse que l'imputation selon la règle choisie cause peu ou pas d'erreur systématique dans les estimations. La variance de l'estimation d'un total est accrue par l'imputation, parce que l'imputation (sauf dans des circonstances tout à fait exceptionnelles) ne reproduit pas la vraie valeur y_k . On le constate concrètement quand on applique la règle d'imputation à des unités de l'échantillon réellement observées; des erreurs se produisent toujours. Si la règle n'est pas exempte d'erreur pour les unités répondantes, elle n'est pas non plus exempte d'erreur pour les unités non répondantes. À la section 4, nous exprimons la variance de t_{\star} comme la somme de deux composantes, une variance d'échantillonnage et une variance attribuable à l'imputation:

$$V^{\text{tot}} = V^{\text{sam}} + V^{\text{imp}}.$$

La variance d'imputation V^{imp} est nulle si toutes les données sont réellement des valeurs observées, ou si la méthode d'imputation est apte à reproduire exactement la vraie valeur y_k pour chaque unité exigeant une imputation. (Aucun de ces deux cas n'est susceptible de survenir en pratique.) Le processus décrit à la section 4 utilise les données après imputation, $y_{\star k}$, $k \in s$, pour obtenir des estimations de chacune des deux composantes, ce qui donne:

$$V^{\text{tot}} = V^{\text{sam}} + V^{\text{imp}}.$$

La composante V^{sam} est déterminée en deux étapes:

- 1) Calculer l'estimation ordinaire, fondée sur le plan, de la variance au moyen des données après imputation. (Par exemple, si l'EASSR est utilisé et que $r = s$, l'estimation non biaisée ordinaire de la variance de $N \bar{y}_s$ est $N^2 (1/n - 1/N) \sum s (y_k - \bar{y}_s)^2 / (n - 1)$. Cette formule, appliquée aux données après imputation, donne $N^2 (1/n - 1/N) \sum s (y_{\star k} - \bar{y}_{\star s})^2 / (n - 1)$, où $\bar{y}_{\star s}$ est la moyenne des n valeurs $y_{\star s}$.)

L'imputation multiple est attrayante du fait qu'elle incorpore l'idée d'une variabilité dans l'imputation. C'est précisément cette variabilité – la variabilité à l'intérieur des ensembles complets de données et entre ces ensembles – qui est exploitée dans les méthodes d'estimation de la variance liées à l'imputation multiple. Ces méthodes font un usage très fructueux des notions statistiques de base. (On pourrait faire valoir, toutefois, que la sélection d'un échantillon comporte elle aussi une variabilité, mais que la plupart des enquêtes doivent se limiter au prélèvement d'un seul échantillon et que l'estimation doit se faire uniquement à partir de ce dernier.)

On peut constater, à l'aide d'exemples simples, que le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une grave sous-estimation de l'incertitude réelle; les responsables de l'échantillonnage sont depuis longtemps sensibles à ce danger. Et il est vrai que parfois, les utilisateurs traitent les valeurs imputées exactement comme les valeurs observées, avec comme résultat des énoncés de précision erronés. Avec les moyens informatiques d'aujourd'hui, il est facile de faire une imputation selon une règle ou une autre, mais il est plus difficile d'obtenir des estimations valides de la variance.

La citation ci-dessus laisse entendre que l'imputation d'une valeur unique, parce qu'aucune variation ne lui est associée, ne permet pas d'en arriver à des estimations raisonnables de la variance, et conduit nécessairement à une sous-estimation. Je ne partage pas cet avis. Les méthodes que j'examine montrent que des estimations valides de la variance sont bel et bien possibles avec l'imputation simple.

Une méthode d'estimation de la variance en présence de valeurs imputées devrait avoir les propriétés suivantes: a) reposer sur un fondement théorique solide; b) être robuste à l'égard des hypothèses régissant l'imputation; c) être pratique, simple à réaliser et facilement acceptée par les utilisateurs.

Bien que l'imputation multiple ait les qualités énoncées en a) et b), il est clair qu'elle est dépourvue de la propriété c), à tout le moins dans certaines applications. Dans l'élaboration du SGE, nous devons miser sur des méthodes d'application facile, après avoir recueilli facilement l'adhésion de l'utilisateur. L'utilisateur de l'ensemble de données (quelqu'un qui n'est pas avant tout un statisticien) peut comprendre facilement que le statisticien fait une imputation unique, avec l'objectif d'attribuer la meilleure valeur possible à l'élément manquant. Certes, il peut être intéressant à certaines fins, par exemple pour des analyses secondaires, de disposer de plusieurs ensembles complets de données, mais les coûts de stockage de tels ensembles obligeront souvent à renoncer à une telle option.

Il est fort possible que l'imputation multiple se révèle utile dans des contextes différents, et pour des raisons autres que celles qui sont essentielles à l'élaboration du SGE. L'imputation multiple est l'un des moyens qui existent pour traiter le problème de la sous-estimation de la variance, du moins dans certaines situations. Cette méthode, par ailleurs récemment critiquée par Fay (1991), ne constitue pas la seule solution possible. Voyons les possibilités offertes par les méthodes d'imputation simple. Le méthode décrite ci-dessous se fonde sur Sarnadal (1990).

3. VARIANCE D'IMPUTATION ET VARIANCE D'ÉCHANTILLONNAGE

Une règle d'imputation correspond à un modèle (explicite ou implicite) de la relation entre les variables d'intérêt de l'enquête. Autrement dit, dès que l'analyste a établi une règle d'imputation, il ou elle a, en fait, choisi un modèle. Le principe du développement qui suit est que si une telle règle est considérée suffisamment bonne pour les estimations ponctuelles (absence d'erreur systématique), elle est également suffisamment bonne pour les estimations connexes de la variance. En d'autres termes, le modèle devrait permettre à la fois le contrôle du biais et la validité de l'estimation de la variance.

de la qualité, de telle sorte que pratiquement toute règle raisonnable . . . d'imputation donnera grosso modo les mêmes résultats . . . Pour ce qui est de l'imputation dans les recensements et les enquêtes par sondage, nous avons adopté une norme selon laquelle un faible niveau d'imputation, de l'ordre de 1 ou 2 %, est ce que nous visons."

Idéalement, un niveau d'imputation de 1 ou 2 % demeure un objectif que nous devrions chercher à atteindre. On note toutefois un taux d'imputation beaucoup plus élevé dans la plupart des enquêtes réalisées de nos jours par les grands organismes d'enquête. Il est clair que si 30 % des valeurs sont imputées, les effets de l'imputation ne peuvent être passés sous silence. L'imputation peut engendrer une erreur systématique (biais) de l'estimation ponctuelle; c'est peut-être là son plus grand danger. Toutefois, même si l'on peut trouver une méthode d'imputation ne produisant pas d'erreur systématique appréciable, on ne peut omettre l'effet souvent considérable qu'exerce l'imputation sur la précision (la variance) de l'estimation ponctuelle. Il importe de se doter de méthodes simples, mais valables, d'estimation de la variance pour les données d'enquête contenant des imputations, afin de pouvoir évaluer adéquatement les coefficients de variation des estimations de l'enquête.

Plusieurs méthodes d'imputation différentes ont été proposées. Celles-ci peuvent être classées de diverses façons. L'un des classements possibles fait intervenir le nombre d'imputations effectuées. Dans les méthodes d'**imputation simple**, une seule valeur est imputée en remplacement d'une donnée manquante. On obtient ainsi un tableau de données complet, dans lequel les valeurs imputées sont marquées. Les estimations sont calculées au moyen de l'ensemble de données complet. Dans le cas de l'**imputation multiple**, deux valeurs ou plus sont imputées à chaque valeur manquante. Plusieurs ensembles de données complets sont ainsi obtenus. Les estimations sont calculées au moyen des ensembles de données complets.

Les méthodes d'imputation diffèrent également selon le modèle qui sous-tend l'imputation. Certaines méthodes se fondent sur un modèle **explicite**, par exemple la valeur imputée est obtenue par un ajustement de régression, un quotient ou une moyenne. Dans d'autres méthodes, le modèle est seulement **implicite**, par exemple dans l'imputation hot deck et l'imputation du plus proche voisin. Les distinctions qui précèdent sont importantes dans le cadre du présent article.

Statistique Canada utilise actuellement diverses méthodes d'imputation: plus proche voisin, quotient courant, moyenne courante, valeur précédente, moyenne précédente, tendance auxiliaire. Ce sont toutes des méthodes d'imputation simple. Les valeurs imputées sont produites dans le Système généralisé de vérification et d'imputation (SGVI); de là elles entrent dans le Système généralisé d'estimation (SGE), où les estimations ponctuelles et les estimations de la variance sont calculées dans plusieurs modules d'estimation différents. Le présent article traite en particulier de la méthode d'imputation par le quotient courant, l'une de celles fondées sur un modèle explicite.

2. QUELQUES RÉFLEXIONS SUR L'IMPUTATION MULTIPLE

La méthode de l'imputation multiple a été proposée par D.B. Rubin vers 1977. Ses idées sont traitées dans plusieurs articles, en particulier dans Herzog et Rubin (1983) et Rubin (1986) qui en exposent la teneur, ainsi que dans un livre, Rubin (1987). L'imputation multiple présente à la fois des avantages et des inconvénients; il en va de même de l'imputation simple.

Rubin (1986) voit dans l'imputation simple l'inconvénient suivant: "... la valeur imputée exclut toute incertitude sur la valeur à imputer. Si une valeur était vraiment appropriée, elle ne serait pas manquante. Ainsi, lorsqu'on assimile les valeurs imputées à des valeurs observées, on sous-estime systématiquement l'élément d'incertitude même en supposant que l'on connaisse les motifs exacts de la non-réponse."

Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation

CARL-ERIK SÄRNDAHL¹

RÉSUMÉ

Dans presque toutes les grandes enquêtes, l'imputation est utilisée sous une forme ou sous une autre. Le présent article expose une méthode d'estimation de la variance lorsque l'imputation simple (plutôt que multiple) est utilisée pour créer un ensemble complet de données. L'imputation ne reproduit jamais les vraies valeurs (sauf dans des cas tout à fait exceptionnels). L'erreur totale de l'estimation de l'enquête est considérée, dans cet article, comme la somme de l'erreur d'échantillonnage et de l'erreur d'imputation. Par conséquent, on en arrive à une variance globale qui est la somme d'une variance d'échantillonnage et d'une variance d'imputation. L'article s'attarde principalement à l'estimation de ces deux composantes, en utilisant les données après imputation, c'est-à-dire les valeurs réellement observées et les valeurs imputées. La méthode est fondée sur un modèle, en ce sens que le modèle sous-jacent à la méthode d'imputation, ainsi que la distribution de randomisation ayant servi à la sélection de l'échantillon, détermineront ensemble la forme que prendront les estimateurs de la variance. Les résultats théoriques sont confirmés par une simulation de Monte Carlo.

MOTS CLÉS: Imputation simple; estimation de la variance; modèle d'imputation; inférence fondée sur un modèle.

1. DIFFÉRENTS TYPES D'IMPUTATION

Le présent article fait état de travaux réalisés dans le cadre de l'élaboration du Système généralisé d'estimation (SGE) de Statistique Canada. Des estimations de la variance seront régulièrement effectuées à l'aide des différents modules d'estimation définissant le SGE. Il est apparu nécessaire de mettre au point des méthodes qui puissent servir à l'estimation de la variance dans les cas où l'ensemble de données contient des valeurs imputées, comme c'est le cas dans la plupart des enquêtes.

La pondération et l'imputation sont deux des principales méthodes permettant de traiter les données manquantes à des fins d'estimation. Dans les publications récentes traitant de la pondération, on considère habituellement que les poids servant à compenser la non-réponse devraient être l'inverse des probabilités de réponse propres à un mécanisme de réponse dont on fait l'hypothèse. Puisque les probabilités de réponse, en général, sont inconnues, elles doivent être estimées à partir des données disponibles. L'imputation, par contre, a l'avantage de fournir un tableau de données complet. Un tel tableau simplifie le traitement des données, mais cela ne signifie pas pour autant que les "méthodes d'estimation ordinaires" peuvent être utilisées directement. Les valeurs imputées se fondent sur des échantillons, et ont par conséquent leurs paramètres statistiques propres, comme une moyenne et une variance.

De nos jours, l'imputation est une technique abondamment utilisée. Il est intéressant de signaler ce que Pritzker, Ogus et Hansen (1965) disent de la ligne de conduite adoptée en matière d'imputation par le US Bureau of the Census: "Essentiellement, notre approche en ce qui concerne le problème de... l'imputation est que nous devrions recueillir de l'information par mesure directe pour une proportion très élevée des agrégats à totaliser, avec un contrôle suffisant

¹ Carl-Erik Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec) H3C 3J7.

- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- NATHAN, G., et HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B*, 42, 377-386.
- PATIL, G.P., et RAO, C.R. (1978). Weighted distributions and size biased sampling with application to wildlife populations and human families. *Biometrics*, 34, 179-189.
- PFEFFERMANN, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association*, 83, 824-833.
- PFEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *Revue Internationale de Statistique* (À paraître).
- RAO, C.R. (1965). On discrete distributions arising out of methods of ascertainment. Dans *Classical and Contagious Discrete Distributions*, (Ed. G.P. Patil). Calcutta: Statistical Publishing Society, 320-332.
- RAO, C.R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? Dans *A Celebration in Statistics* (Eds. A.C. Atkinson et S.E. Fienberg). New York: Springer-Verlag, 543-569.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 469-474.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.
- RUBIN, D.B. (1985). The use of propensity scores in applied Bayesian inference. Dans *Bayesian Statistics 2*, (Eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley et A.F.M. Smith). Amsterdam: Elsevier Science, 463-472.
- SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.
- SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10, 616-620.

pour le recensement. Les estimateurs tirés de distributions pondérées sont particulièrement efficaces dans l'étude de simulation en présence d'un plan d'échantillonnage informatif, ce qui n'est pas le cas de l'EMV "exact", qui peut devenir fortement biaisé en présence d'un tel plan. Toutefois, pour pouvoir utiliser ces estimateurs, il faut modéliser la relation entre les probabilités d'échantillonnage et les données de l'échantillon. L'étude de simulation montre qu'une modélisation ou une estimation erronées de la relation peuvent créer des biais considérables.

Par conséquent, la question qu'il faut se poser par rapport à l'usage de ces estimateurs est de savoir si les données de l'échantillon peuvent reproduire fidèlement le modèle qui décrit la relation entre les probabilités d'échantillonnage d'une part et les variables de réponse et les variables de plan d'autre part. Il semble que la seule manière de répondre à cette question soit d'examiner des enquêtes réelles qui utilisent des plans d'échantillonnage communs. Notons d'autres questions importantes qui ont trait à l'utilisation de ces estimateurs, à savoir i) l'existence d'estimateurs de variance fiables pour permettre la construction d'intervalles de confiance justes et ii) la manière de parer à une erreur de spécification de la distribution des variables de réponse pour la population. Ces deux questions se posent aussi bien pour d'autre méthodes d'estimation fondées sur le principe du maximum de vraisemblance. Nous espérons que les résultats de notre étude ouvriront la voie à d'autres recherches visant à approfondir ces questions et des questions connexes.

REMERCIEMENTS

Danny Pfeffermann est bénéficiaire du Programme de bourses de recherche et de stages de Statistique Canada.

BIBLIOGRAPHIE

- ANDERSON, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52, 200-203.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- CHAMBERS, R.L. (1986). Design adjusted parameter estimation. *Journal of the Royal Statistical Society A*, 149, 161-173.
- CHAMBLESS, L.E., et BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.
- COX, D.R. (1969). Some sampling problems in technology. Dans *New Developments in Survey Sampling*, (Eds. N. Johnson et H. Smith Jr.), New York: Wiley, 506-527.
- GODAMBE, V.P., et RAJARSHI, M.B. (1989). Optimal estimation for weighted distributions: semi-parametric models. Dans *Statistical Data Analysis and Inference*, (Ed. Y. Dodge), Amsterdam: Elsevier Science, 199-208.
- GODAMBE, V.P., et THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- HAUSMAN, J.A., et WISE, D.A. (1981). Stratification on endogenous variables and estimation; the Gary Income Maintenance Experiment. Dans *Structure Analysis of Discrete Data with Econometric Applications*, (Eds. C.F. Mansky et D. McFadden). Cambridge, Mass.: MIT Press, 366-391.
- HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A*, 143, 474-487.

Tableau 3

EQM des estimateurs de B_{Z1} pour des plans d'échantillonnage différents et des variables de plan différentes (coefficient vrai: $B_{Z1} = 1.33$)

Estimateurs	D1 - échantillonnage avec PPT		D2 - échantillonnage stratifié	
	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$
$ML(Z_1, Z_2)$	0.043	0.069*	0.048	0.120*
$WML(Z_1, Z_2)$	0.054	0.060	0.068	0.066
$ML(Z_1)$	0.045	0.078*	0.056	0.134*
$WML(Z_1)$	0.055	0.062	0.069	0.065
$WDMML(X^*, Z_1)$	0.043	0.047	0.049	0.045
$WDMML(X^*)$	0.044	0.049	0.050	0.046
CPL	0.055	0.063	0.069	0.065
$WDMML(X^*, \tilde{t}_{LS})$	-	-	0.048	0.045
$WDMML(X^*, \tilde{z}_{LS}, Z_1)$	-	-	0.048	0.045

* EQM expliquée en très grande partie par le biais.

logistique. Cette dernière solution a ceci d'intéressant qu'elle permet de réduire le nombre de paramètres par rapport auxquels la fonction de vraisemblance doit être maximisée, ce qui peut être indispensable lorsqu'on a un nombre élevé de strates.

Nous avons examiné jusqu'ici le biais et l'EQM inconditionnels des estimateurs. Comme nous le mentionnions dans la section 4.1, nous avons aussi étudié les propriétés conditionnelles de ces estimateurs en calculant le biais et l'EQM pour des échantillons qui ont des moyennes d'échantillon de la variable de plan semblables. Les conclusions de cette étude rejoignent celles que nous avons exposées plus tôt. Par conséquent, des estimateurs qui sont approximativement non biaisés inconditionnellement sont aussi approximativement non biaisés conditionnellement et vice versa.

Cette observation est quelque peu étonnante car on a souvent mis en évidence dans les ouvrages statistiques les piètres propriétés conditionnelles de l'estimateur CPL par exemple. Cette contradiction s'explique peut-être par le fait que la taille de nos échantillons est élevée ou que le nombre de groupes (10) qui a servi à la répartition des échantillons n'est pas suffisamment grand. Pour des considérations d'espace, nous ne reproduisons pas ici les résultats de l'analyse des propriétés conditionnelles des estimateurs.

5. CONCLUSIONS

Les résultats de l'étude de simulation montrent que les estimateurs que l'on obtient par la maximisation de fonctions de vraisemblance fondées sur des distributions pondérées peuvent remplacer avantageusement les pseudo-estimateurs du maximum de vraisemblance que l'on obtient en maximisant les estimateurs convergents selon le plan des équations de vraisemblance

4.4 Résultats

Nous considérons que les résultats qui découlent de l'estimation de $\mu_1 = E(Y_1)$, de $\sigma_1^2 = \text{Var}(Y_1)$, et de B_{21} – le coefficient de pente de la régression de Y_2 par rapport à Y_1 – représentent bien les résultats que l'on obtient en estimant les autres paramètres. Les tableaux 1 à 3 contiennent l'EQM des divers estimateurs pour deux plans d'échantillonnage différents et deux valeurs de la variable de plan différentes. Les EQM qui sont expliquées en très grande partie par le biais sont identifiées par un astérisque.

On peut résumer ainsi les principales constatations qui ressortent de ces tableaux (et de l'estimation des autres paramètres du modèle):

- 1) L'estimateur $ML(Z_1, Z_2)$ surclasse tous les autres lorsque les conditions de "neutralité" sont respectées, mais il est fortement biaisé lorsque le plan d'échantillonnage est informatif. L'estimateur $WML(Z_1, Z_2)$ est essentiellement non biaisé dans tous les cas, sauf que l'utilisation de poids d'échantillonnage accroît la variance. Néanmoins, cet estimateur surclasse en général l'estimateur CPL, surtout en ce qui concerne l'échantillonnage avec PPT à cause de l'utilisation des valeurs de (Z_1, Z_2) pour la population.
- 2) L'estimateur $ML(Z_1)$ est fortement biaisé dans presque tous les cas. Notons particulièrement le fort biais observé lorsque $t_i = 0.5(z_{1i} + z_{2i})$, ce qui témoigne de la sensibilité des EMV au regard de la spécification exacte des variables de plan. Comme $WML(Z_1, Z_2)$, l'estimateur $WML(Z_1)$ est non biaisé et surclasse l'estimateur CPL en ce qui a trait à l'échantillonnage avec PPT.
- 3) L'estimateur CPL est non biaisé dans tous les cas. Une constatation intéressante ressort des tableaux à propos de cet estimateur; en effet, par rapport aux autres estimateurs étudiés, le CPL est plus efficace dans l'estimation de la moyenne que dans l'estimation des variances et des covariances. Intuitivement, nous pouvons expliquer cette différence par le fait que dans l'estimation des variances et covariances, les poids d'échantillonnage sont utilisés deux fois, ce qui a pour effet d'accroître la variance.

- 4) En ce qui concerne l'échantillonnage avec PPT, les estimateurs $WDM(X^*)$ et $WDM(X^*, Z_1)$ présentent un très bon rendement, le premier surclassant nettement CPL et le second l'emportant sur $WML(Z_1)$. Fait intéressant, $WDM(X^*)$ est généralement plus efficace que $WML(Z_1)$ même s'il utilise moins d'information. On peut expliquer la supériorité de $WDM(X^*)$ sur CPL par le fait que le premier repose plus souvent sur des modèles, bien que, comme il en a été question dans la section 2.4, on puisse imaginer CPL comme l'estimateur qui maximise l'estimateur non biaisé selon le plan des équations de vraisemblance applicables à la population.
- 5) Voyons maintenant l'échantillonnage stratifié. Dans le cas d'un plan informatif, c'est-à-dire lorsque $t_i = 0.25z_i$, la situation ressemble à celle observée pour l'échantillonnage avec PPT: $WDM(X^*)$ surclasse une fois de plus CPL et $WML(Z_1)$. En fait, aucun des quatre estimateurs issus de fonctions de vraisemblance fondées sur des distributions pondérées ne se distingue vraiment des autres malgré que les quatre utilisent des données d'échantillon et de population différentes. Lorsque $t_i = 0.5z_i$, les quatre mêmes estimateurs sont inférieurs à $WML(Z_1)$ et à CPL sauf – fait très intéressant – en ce qui a trait à l'estimation du coefficient de régression, où ils sont à peu de choses près aussi efficaces que l'estimateur optimal $ML(Z_1, Z_2)$. Le rendement médiocre de $WDM(X^*)$ (et, dans une beaucoup moins grande mesure, de $WDM(X^*, Z_1)$) dans l'estimation de la moyenne et de la variance est principalement attribuable à une mauvaise définition des limites de strates et, par conséquent, à une spécification inexacte du dénominateur de l'expression (3.10). On peut probablement résoudre la difficulté soit en incluant dans l'ensemble des paramètres inconnus de la fonction de vraisemblance (3.10) les limites de strates et les coefficients α^* qui mettent en relation les valeurs t_i et les observations (équation 3.7), soit en remplaçant la fonction discriminante linéaire par une autre fonction (non linéaire) comme la régression

Tableau 1

EQM des estimateurs de μ_1 pour des plans d'échantillonnage différents et des variables de plan différentes (moyenne vraie: $\mu_1 = 40$)

Estimateurs	D1 - échantillonnage avec PPT		D2 - échantillonnage stratifié	
	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$
$ML(Z_1, Z_2)$	0.43	1.86*	0.47	3.43*
$WML(Z_1, Z_2)$	0.43	0.57	0.50	0.52
$ML(Z_1)$	2.67*	4.38*	6.39*	8.32*
$WML(Z_1)$	0.58	0.90	0.62	0.58
$WDM L(X^*, Z_1)$	0.56	0.63	1.51*	0.59
$WDM L(X^*)$	0.80	0.90	3.59*	0.49
CPL	0.77	1.19	0.56	0.47
$WDM L(X^*, \tilde{t}_{\tilde{z}})$	—	—	0.74	0.43
$WDM L(X^*, \tilde{t}_{\tilde{z}}, Z_1)$	—	—	0.74	0.57

* EQM expliquée en très grande partie par le biais.

Tableau 2

EQM des estimateurs de σ_1^2 pour des plans d'échantillonnage différents et des variables de plan différentes (variance vraie: $\sigma_1^2 = 300$)

Estimateurs	D1 - échantillonnage avec PPT		D2 - échantillonnage stratifié	
	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$	$t_l = 0.5\tilde{z}_l$	$t_l = 0.25\tilde{x}_l$
$ML(Z_1, Z_2)$	12.33	18.35*	16.00	29.00*
$WML(Z_1, Z_2)$	14.00	18.72	20.87	19.83
$ML(Z_1)$	24.32*	33.66*	35.16*	53.66*
$WML(Z_1)$	18.61	26.61	24.22	20.35
$WDM L(X^*, Z_1)$	14.36	17.41	26.94*	15.49
$WDM L(X^*)$	16.37	19.68	41.08*	15.34
CPL	21.11	29.06	24.19	20.18
$WDM L(X^*, \tilde{t}_{\tilde{z}})$	—	—	26.18*	15.46
$WDM L(X^*, \tilde{t}_{\tilde{z}}, Z_1)$	—	—	25.70*	15.72

* EQM expliquée en très grande partie par le biais.

PLAN D2

Les cinq premiers estimateurs sont les mêmes que pour le plan D1. Les quatre autres sont définis ci-dessous:

$WDM L(X^*)$ – Les estimateurs que l'on obtient par la maximisation de la fonction de vraisemblance (3.10), les coefficients α^* [(équation (3.7)] étant estimés au moyen de la fonction discriminante linéaire.

$WDM L(X^*, Z_1)$ – Même chose que $WDM L(X^*)$, sauf que la moyenne et la variance de Z_1 sont posées égales à leurs valeurs respectives pour la population.

$WDM L(X^*, \tilde{t}_s)$ – Les estimateurs que l'on obtient par la maximisation de la fonction de vraisemblance (3.10) lorsque les valeurs $\tilde{t}_s = (t_1, \dots, t_n)$ sont connues pour les unités de l'échantillon.

$WDM L(X^*, \tilde{t}_s, Z_1)$ – Même chose que $WDM L(X^*, \tilde{t}_s)$, sauf que la moyenne et la variance de Z_1 sont posées égales à leurs valeurs respectives pour la population.

Il convient de souligner que les estimateurs tirés de distributions pondérées ne sont pas de vrais EMV en raison, comme l'explique la section 3.2, des approximations incluses dans le processus de maximisation (voir aussi le commentaire 2 ci-dessous).

Commentaires

1) Les estimateurs que nous étudions peuvent être classés en fonction des données d'échantillon et de population qu'ils utilisent et selon que les variables de plan sont spécifiées correctement ou non et que les conditions de "neutralité" sont respectées ou non. Ainsi, les estimateurs $ML(Z_1, Z_2)$ et $WML(Z_1, Z_2)$ utilisent les valeurs de Z_1 et Z_2 pour la population ainsi que les valeurs de Y_1 et Y_2 pour l'échantillon. Comme nous l'avons mentionné dans la section 2.4 et comme le souligne aussi Pfeffermann (1992) dans une analyse plus poussée, l'emploi de $WML(Z_1, Z_2)$ a pour but de parer à d'éventuelles erreurs de spécification du modèle ou à des plans d'échantillonnage informatifs. Les estimateurs $ML(Z_1)$, $WML(Z_1)$, $WDM L(X^*, Z_1)$ et $WDM L(X^*, \tilde{t}_s, Z_1)$, utilisent les valeurs connues de Z_1 pour la population mais non les valeurs de Z_2 , pas même celles pour l'échantillon. On se sert de ces estimateurs lorsque les variables de plan sont mal spécifiées ou que la valeur de certaines d'entre elles n'est pas connue. L'estimateur $WDM L(X^*)$ utilise uniquement les valeurs de Y_1 , Y_2 et Z_1 relatives à l'échantillon ainsi que les probabilités d'échantillonnage. L'estimateur $WDM L(X^*, \tilde{t}_s)$ utilise, outre ces valeurs et ces probabilités, les valeurs d'échantillonnage de la variable de plan. L'estimateur CPL, quant à lui, utilise seulement les valeurs de Y_1 et Y_2 pour l'échantillon ainsi que les probabilités d'échantillonnage.

2) Nous avons maximisé la fonction de vraisemblance tirée de distributions pondérées en nous servant d'une quasi-méthode de Newton puisée dans la bibliothèque de sous-programmes d'IMS L. Cette méthode ne peut être appliquée que si l'utilisateur a fourni au préalable les dérivées partielles de la fonction de vraisemblance par rapport à chacun des paramètres. Il convient ici de souligner une observation qui a été faite durant le processus de maximisation. Il est plus facile de paramétrer la fonction de vraisemblance par rapport à Σ^{-1} , où Σ est la matrice de covariance de Y_1 , Y_2 et Z_1 . De plus, pour faire en sorte que les six paramètres qui définissent Σ^{-1} ne soient assujettis à aucune contrainte, nous utilisons les éléments de la matrice triangulaire supérieure B de sorte que $B'B = \Sigma^{-1}$. Peu importe les valeurs que nous aurons pour B , nous obtiendrons toujours une matrice Σ^{-1} semi-définie positive.

multidimensionnelle. Dans la seconde étape, nous avons prélevé des échantillons indépendants de taille $n = 800$ en nous servant des deux plans d'échantillonnage décrits dans la section 3.2 et de deux définitions différentes pour la variable de plan. Le nombre d'échantillons prélevés dans chaque cas était de 300. Nous avons calculé les divers estimateurs pour chacun des échantillons sur la base des données empiriques disponibles, puis nous avons calculé le biais et l'erreur quadratique moyenne (EQM) empiriques par rapport à l'ensemble des échantillons. Afin d'étudier et de comparer les propriétés conditionnelles des estimateurs analysés, nous avons divisé les 300 échantillons prélevés dans chaque cas en 10 groupes selon l'ordre croissant des valeurs de la moyenne d'échantillon de la variable de plan et nous avons calculé le biais et l'EQM pour chacun des groupes. Nous décrivons plus en détail ci-dessous les diverses étapes de la simulation.

4.2 Génération des valeurs de la population et plans d'échantillonnage

Les valeurs de z_{1i} et de z_{2i} ont été générées de façon indépendante à partir d'une distribution normale $(20, 10^2)$. Les valeurs de y_{1i} , quant à elles, ont été générées comme $y_{1i} = z_{1i} + z_{2i} + \epsilon_{1i}$; $\epsilon_{1i} \sim N(0, 10^2)$. Enfin, les valeurs y_{2i} ont été générées comme $y_{2i} = y_{1i} + 0.5z_{1i} + 0.5z_{2i} + \epsilon_{2i}$; $\epsilon_{2i} \sim N(0, 20^2)$.

Nous nous sommes servis des deux plans d'échantillonnage décrits dans la section 3.2 et de deux définitions différentes pour la variable de plan (en l'occurrence, la taille): i) $t_i = 0.5(z_{1i} + z_{2i})$ et ii) $t_i = 0.25(y_{1i} + y_{2i} + z_{1i} + z_{2i})$. En conséquence, un échantillonnage fondé sur la première variable de plan respecte les conditions de "neutralité" définies dans la section 2.1, pourvu que l'on connaisse les valeurs (Z_1, Z_2) pour la population entière. Si ces valeurs sont connues uniquement pour l'échantillon, le plan d'échantillonnage ne sera "neutre" que par rapport à la distribution conditionnelle $f(y_1, y_2 | z_1, z_2)$. Lorsque l'échantillonnage repose sur la seconde variable de plan, nous avons un plan d'échantillonnage informatif. En ce qui concerne l'échantillonnage stratifié (plan D2), nous avons généré huit strates de taille identique suivant l'ordre croissant des valeurs de la variable de taille. Les effectifs des strates sont définis de telle manière qu'ils augmentent avec la valeur de t_i/s .

4.3 Estimateurs étudiés

Les paramètres estimés dans notre étude sont le vecteur de moyennes et la matrice de variances-covariances de la distribution marginale de (Y_1, Y_2) . Nous allons étudier sept estimateurs pour le plan D1 et neuf pour le plan D2. Le lecteur est prié de se référer à la section 3.2 pour une description du mode de calcul des divers estimateurs.

PLAN D1

- $ML(Z_1, Z_2)$ – L'EMV exact lorsque le plan est "neutre" (équation 2.4).
- $WML(Z_1, Z_2)$ – Les estimateurs que l'on tire de $ML(Z_1, Z_2)$ en remplaçant les statistiques non pondérées par des statistiques pondérées qui suivent l'équation (2.7)).
- $ML(Z_1)$ – Même chose que $ML(Z_1, Z_2)$ sauf que Z_1 est la seule variable de plan de sorte que $\tilde{Z} = Z_1$.
- $WML(Z_1)$ – Même chose que $WML(Z_1, Z_2)$ sauf que Z_1 est la seule variable de plan.
- CPL – Les pseudo-estimateurs du maximum de vraisemblance classiques (équation 2.7).
- $WDML(X^*)$ – Les estimateurs (tirés de distributions pondérées) que l'on obtient par la maximisation de la fonction de vraisemblance définie en (3.6).
- $WDML(X^*, Z_1)$ – Les estimateurs que l'on obtient par la maximisation de la fonction de vraisemblance (3.6), sauf que la moyenne et la variance de Z_1 sont posées égales à leurs valeurs respectives pour la population.

(3.9)
$$w = P(i \in s) \approx P_1 \int_{t^{(1)}}^{-\infty} \phi(t) dt + \sum_{h=2}^L P_h \int_{t^{(h-1)}}^{t^{(h)}} \phi(t) dt + P_L \int_{\infty}^{t^{(L-1)}} \phi(t) dt,$$

où $\phi(t)$ est la densité de probabilité normale de T .
Supposons que les strates soient assez grandes pour que l'on puisse considérer les tirages dans chaque strate comme indépendants. Posons $\mu_T = E(T) = \tilde{\alpha} \mu_X, \sigma_T^2 = \tilde{\alpha} \mu_X, \sigma_T^2 = \tilde{\alpha} \sum_{XX} \tilde{\alpha}$ ainsi que $\Phi_h = \int_{t^{(h)}}^{-\infty} \phi(t) dt$. Pour des limites données $\{t^{(h)}\}$ et les coefficients vectoriels $\tilde{\alpha}$, on peut exprimer la fonction de vraisemblance pour \tilde{g} de la façon suivante:

$$L(\tilde{g}; X, s) = \text{const} \times \Pi_{h=1}^n h(\tilde{x}_i; \tilde{g}) \Pi_{L=1}^n P_h^{n_h} /$$

(3.10)
$$\{P_1 \Phi_1 + \sum_{L=2}^L P_h [\Phi_h - \Phi_{h-1}] + P_L [1 - \Phi_{L-1}]\}^n.$$

Hausman et Wise (1981) recourent à une variante de (3.10) pour estimer le vecteur des coefficients de régression lorsque les limites de strates sont déterminées par les valeurs de la variable dépendante. Ils supposent que les limites de strates sont connues mais n'empêchent pas que les probabilités d'échantillonnage dans les strates soient inconnues, auquel cas ces probabilités sont ajoutées à l'ensemble de paramètres inconnus par rapport auxquels la fonction de vraisemblance est maximisée.

Dans beaucoup de cas, les limites de strates sont inconnues et doivent être estimées à l'aide des données de l'échantillon. Lorsque les valeurs $\{t_i, i = 1, \dots, n\}$ sont incluses dans les données, on peut estimer le vecteur $\tilde{\alpha}$ grâce à une régression de t_i par rapport à \tilde{x}_i , comme dans l'exemple de l'échantillonnage PPT ci-dessus. En outre, si $t^{(1)} \leq \dots \leq t^{(n)}$ sont les valeurs ordonnées de t_i , on peut estimer les limites de strates par la formule $t^{(1)} = 1/2(t^{(n_1)} + t^{(n_1+1)}) \dots t^{(L-1)} = 1/2(t^{(n^*)} + t^{(n^*+1)})$, où $n^* = \sum_{L=1}^{L-1} n_h$. En substituant ces estimations dans l'équation (3.10), on obtient une version approchée de la fonction de vraisemblance qui peut ensuite être maximisée par rapport à \tilde{g} .

La situation se complique lorsqu'on ne connaît pas les valeurs t_i même pour les unités de l'échantillon. Nous tentons de résoudre cette difficulté dans l'étude de simulation en prévoyant la valeur de t_i au moyen de la fonction discriminante linéaire de Fisher, c'est-à-dire en spécifiant les coefficients vectoriels $\tilde{\alpha}$ de telle manière qu'ils maximisent le rapport de la somme des carrés intergroupe à la somme des carrés intra-groupe des combinaisons linéaires $\tilde{\alpha}' X_i$, les groupes étant ici les strates. Une fois constitués les prédicteurs $\hat{t}_i = \tilde{\alpha}' \tilde{x}_i$, on estime les limites de strates de la même manière que dans le cas précédent sauf que t_i remplace t_i . De plus, $\hat{\mu}_T = \tilde{\alpha}' \mu_X$ et $\hat{\sigma}_T^2 = \tilde{\alpha}' \sum_{XX} \tilde{\alpha}$. En substituant ces estimateurs dans l'équation (3.10), on obtient une version approchée de la fonction de vraisemblance qui peut ensuite être maximisée par rapport à \tilde{g} .
Comme dans l'exemple de l'échantillonnage PPT, il est possible d'adapter la fonction de vraisemblance (3.10) à la situation où seulement quelques-unes des variables de plan seraient connues. La maximisation de la nouvelle fonction s'effectue de la manière décrite plus haut.

4. RÉSULTATS DE LA SIMULATION

4.1 Remarques générales

Afin d'illustrer et de comparer l'efficacité des diverses MMV présentées dans cet article, nous avons exécuté une petite étude de simulation qui se divise en deux étapes. La première étape consistait à générer une population finie de taille $N = 8,000$ de telle manière que les valeurs $\tilde{x}_i' = (y_{1i}, y_{2i}, z_{1i}, z_{2i})', i = 1, \dots, 8,000$ proviennent d'une distribution normale

Lorsque $\tilde{\alpha}_1 = \tilde{0}$ et que T est connu pour chaque unité de la population, on peut estimer les paramètres inconnus à l'aide de l'équation (2.3). Les EMV correspondants figurent en (2.4), où \tilde{Z} est remplacé par T . Supposons, toutefois, que la seule information dont dispose l'analyste consiste dans les valeurs de l'échantillon, $\tilde{x}_i' = (\tilde{y}_i', \tilde{z}_i')$, $i = 1, \dots, n$, et les probabilités d'échantillonnage, $P_i = t_i/NT$. Suivant l'hypothèse $T = \mu_T$, on peut exprimer la fonction de vraisemblance pour $[\tilde{\mu}_X, \Sigma^{XX}]$ de la façon suivante (en se servant de l'équation (3.3))

$$(3.5) \quad L(\tilde{\mu}_X, \Sigma^{XX}; X_{s,s}) = \prod_{i=1}^n (\tilde{\alpha}' \tilde{x}_i' \phi(\tilde{x}_i'; \tilde{\mu}_X, \Sigma^{XX}) / (\tilde{\alpha}' \tilde{\mu}_Y + \tilde{\alpha}_2' \tilde{\mu}_Z)^n,$$

où $\phi(\tilde{x}; \tilde{\mu}_X, \Sigma^{XX})$ est la densité de probabilité normale de moyenne $\tilde{\mu}_X$ et de matrice de variances-covariances Σ^{XX} . La fonction de vraisemblance définie en (3.5) est aussi une fonction des coefficients vectoriels inconnus $\tilde{\alpha}$. Il est toutefois possible de déterminer les valeurs de $\tilde{\alpha}$ jusqu'à concurrence d'une constante c (qui s'annule dans la fonction de vraisemblance) en faisant une régression des probabilités d'échantillonnage P_i par rapport à $\tilde{\alpha}$.

Dans l'étude de simulation présentée dans la section 4, nous envisageons le cas où les variables de plan ne sont pas toutes connues, même pour les unités de l'échantillon. Par conséquent, supposons que $\tilde{Z}_i' = (Z_{1i}, Z_{2i})$ et que l'information dont dispose l'analyste consiste dans les probabilités d'échantillonnage P_i , $i = 1, \dots, n$, et les observations $\{\tilde{x}_i'\} = (\tilde{y}_i', z_{1i})$, $i = 1, \dots, n$. La fonction de vraisemblance (3.3) s'écrit maintenant

$$(3.6) \quad L(\tilde{\mu}_X^*, \Sigma^{XX*}; X_{s,s}^*) = \prod_{i=1}^n w(\tilde{x}_i^*) \phi(\tilde{x}_i^*; \tilde{\mu}_X^*, \Sigma^{XX*}) / (w^*)^n,$$

où $w(\tilde{x}_i^*)$ représente les probabilités d'échantillonnage exprimées comme une fonction de \tilde{x}_i^* . De toute évidence, les valeurs \tilde{x}_i^* ne déterminent pas entièrement les probabilités $w(\tilde{x}_i^*)$, sauf si $\alpha_{22} = 0$. En supposant la normalité,

$$(3.7) \quad w(\tilde{x}_i; \tilde{\alpha}) = \alpha_0^* + \tilde{\alpha}_1^* \tilde{y}_i + \alpha_2^* z_{1i} + \epsilon_i,$$

où $\{\epsilon_i\}$ est du bruit blanc. On peut donc obtenir une version approchée de la fonction de vraisemblance (3.6) en substituant $w^*(\tilde{x}_i^*) = \alpha_0^* + \tilde{\alpha}_1^* \tilde{y}_i + \alpha_2^* z_{1i}$ à $w(\tilde{x}_i^*)$. Les valeurs de $\tilde{\alpha}^* = (\alpha_0^*, \tilde{\alpha}_1^*, \alpha_2^*)'$ peuvent être estimées au moyen de l'équation de régression (3.7), et les estimations introduites dans la fonction de vraisemblance.

D2 - Échantillonnage stratifié où T sert de variable de stratification: Supposons que la population U soit divisée en L strates U_1, \dots, U_L de tailles respectives N_1, \dots, N_L , $\Sigma_{h=1}^L N_h = N$, suivant l'ordre croissant des valeurs de T . Considérons un échantillon aléatoire simple stratifié sans remise de taille $n = \Sigma_{h=1}^L n_h$, où les $\{n_h\}$ sont fixes. La fonction de densité pondérée de X_i^w , les valeurs enregistrées pour l'unité $i \in s$, est en l'occurrence [comparer avec (3.2)]

$$(3.8) \quad h_w(\tilde{x}_i; \tilde{\alpha}, \tilde{\delta}) = f(\tilde{x}_i | i \in s) = \begin{cases} P_1 h(\tilde{x}_i; \tilde{\delta}) / w & \text{si } t_{(1)} \leq t_i \\ P_2 h(\tilde{x}_i; \tilde{\delta}) / w & \text{si } t_{(1)} \leq t_i \leq t_{(2)} \\ \vdots & \vdots \\ P_L h(\tilde{x}_i; \tilde{\delta}) / w & \text{si } t_{(L-1)} \leq t_i \end{cases}$$

où $P_h = (n_h/N_h)$ et où, pour $\{N_h\}$ suffisamment grand, la probabilité $w = P(i \in s)$ peut être reproduite assez fidèlement par l'expression

sorte que la fonction de vraisemblance est

(3.3)
$$L(\hat{\delta}; X, s) = \text{const} \times \prod_{i=1}^n h(\tilde{x}_i; \hat{\delta}) / [\int w(\tilde{x}; \tilde{\omega}) h(\tilde{x}; \hat{\delta}) d\tilde{x}]^n,$$

où $X_i' = [\tilde{x}_1, \dots, \tilde{x}_n]$. La fonction de vraisemblance (3.3) possède les propriétés suivantes:

- 1) Elle est définie en fonction du paramètre vectoriel $\hat{\delta}$, ce qui est un avantage par rapport à la fonction de vraisemblance (2.3), où $\hat{\delta}$ n'est pas directement représenté.

- 2) Elle est une fonction des probabilités d'échantillonnage $w(\tilde{x}_i; \tilde{\omega})$, qui font partie du dénominateur.

- 3) Elle a trait à la distribution conditionnelle des données de l'échantillon étant donné les unités de l'échantillon, ce qui la distingue de la fonction de vraisemblance tirée de la densité de probabilité définie en (2.1), qui est la densité de probabilité conjointe des données de l'échantillon et du vecteur \tilde{I} des indicateurs d'inclusion dans l'échantillon. Godambe et Rajarshi (1989) donnent un exemple de l'utilisation de cette fonction de densité avec des distributions pondérées pour des EMV.

- 4) Pour utiliser la fonction de vraisemblance (3.3), il faut définir la densité de probabilité conjointe, $h(\tilde{x}; \hat{\delta})$, s'appliquant à la population et spécifier la relation entre les probabilités d'échantillonnage et les variables observées pour l'échantillon. La première condition est commune à toutes les méthodes proposées dans les ouvrages statistiques pour l'estimation fondée sur le principe du maximum de vraisemblance. En revanche, la spécification des fonctions $\tilde{w}(\tilde{x})$ est propre à la méthode exposée dans cette section. Cette opération peut se faire par une modélisation de la relation empirique entre les probabilités d'échantillonnage et les observations. Une fois qu'on a trouvé un modèle acceptable, on peut estimer les probabilités $w(\tilde{x}; \tilde{\omega})$ sur la base de l'échantillon, puis intégrer les estimations dans la fonction de vraisemblance. Nous présentons ci-dessous deux exemples que nous analysons empiriquement dans la section 4.

3.2 Exemples

Nous reprenons le modèle qui a été considéré dans la section 2 et dans lequel $\tilde{X}'_i = (\tilde{Y}'_i, \tilde{Z}'_i)$ sont des réalisations indépendantes d'une distribution normale multidimensionnelle de moyenne $\tilde{\mu}'_i = (\tilde{\mu}'_{Y_i}, \tilde{\mu}'_{Z_i})$ et de matrice de variances-covariances.

(3.4)
$$\Sigma_{XX} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}.$$

Considérons les plans d'échantillonnage suivants:

D1 - **Echantillonnage PPT avec remise**: Posons $T_i = \tilde{\alpha}'_i \tilde{Y}_i + \tilde{\alpha}'_2 \tilde{Z}_i$ comme une variable de plan simple et supposons que l'échantillon est prélevé avec des probabilités proportionnelles aux valeurs T de telle sorte qu'à chaque tirage $k = 1, \dots, n$, $P(\text{ies}) = t_i / N T_i$, $i = 1, \dots, N$, où $\bar{T} = \sum_{i=1}^N t_i / N$. Nous supposons que N est suffisamment grand pour que l'écart entre \bar{T} et $\mu_T = E(T)$ soit négligeable. Les coefficients $\tilde{\alpha} = (\tilde{\alpha}'_1, \tilde{\alpha}'_2)$ sont fixes. Dans des cas particuliers, $\tilde{\alpha}'_1 = \tilde{0}$, ce qui implique que T est uniquement fonction des variables de plan (variables auxiliaires) \tilde{Z} , ou $\tilde{\alpha}'_2 = 0$, auquel cas T est uniquement une fonction des variables de réponse \tilde{Y} . Supposons, comme avant, que nous voulons estimer la moyenne μ_Y et la matrice de variances-covariances Σ_{YY} ou des fonctions de ces paramètres.

Les valeurs des variables de plan (variables auxiliaires) Z pour la population peuvent servir à accroître l'efficacité des estimateurs convergents selon le plan de $\tilde{U}(\theta)$. Mentionnons à titre d'exemple l'"EMV pondéré en probabilité" de Nathan et Holt (1980) et de Smith et Holmes (Skinner et coll., 1989, chap. 8). L'estimateur en question a la même structure que l'EMV exact tiré de l'expression (2.4), à la différence que les statistiques non pondérées sont remplacées par des statistiques pondérées. Par exemple, dans l'équation (2.4), $(\tilde{y}_s, \tilde{z}_s)$ est remplacé par $\sum_{i=1}^n w_i^* (\tilde{y}_i, \tilde{z}_i) / \sum_{i=1}^n w_i^*$; de même pour les autres expressions.

Une propriété importante du pseudo-EMV est qu'il est généralement convergent selon le plan pour les valeurs de la population que l'on obtient en résolvant les équations de vraisemblance correspondantes pour le recensement, que le modèle soit exact ou non ou que le plan d'échantillonnage soit informatif ou non. Voir Pfeffermann (1993) en ce qui regarde les conséquences de cette propriété par rapport à d'autres études. Godambe et Thompson (1986) examinent d'autres propriétés théoriques du pseudo-EMV.

3. EMV TIRÉ DE DISTRIBUTIONS PONDÉRÉES

3.1 Formulation générale

La densité de probabilité pondérée d'une variable aléatoire X^w est définie

$$(3.1) \quad f^w(x) = w(x)f(x)/w,$$

où $f(x)$ est la densité de probabilité non pondérée et $w = \int w(x)f(x)dx = E[w(X)]$ est le facteur de normalisation qui fait que la probabilité totale est égale à un. On obtient des conditions propices à la création de distributions pondérées lorsque des réalisations x de $f(x)$ sont observées et enregistrées avec des probabilités différentes, $w(x)$. L'espérance w correspond alors à la probabilité d'enregistrer une observation et $f^w(x)$ est la densité de probabilité de la variable aléatoire correspondante X^w .

La notion de distribution pondérée a été définie par Rao (1965). Patil et Rao (1978) analysent divers cas qui impliquent des densités de probabilité comme celle définie en (3.1). On retrouve dans de nombreuses applications le cas particulier où $w(x) = |x|$, $|x|$ étant une mesure de la grandeur de x . Dans ces circonstances, la densité de probabilité est dite "biaisée par rapport à la grandeur" ou "biaisée par rapport à la valeur". Cox (1969) et Patil et Rao (1978) étudient les propriétés de cette distribution par rapport à diverses fonctions de densité $f(x)$. Vardi (1982), pour sa part, traite l'estimation de distributions pondérées.

Comment la notion de distribution pondérée peut-elle être appliquée à l'inférence analytique fondée sur des échantillons complexes? Considérons, comme avant, une population finie $U = \{1, \dots, N\}$ avec des valeurs aléatoires $X(i) = \tilde{x}_i = (\tilde{y}_i, \tilde{z}_i)$ tirées de façon indépendante d'une densité de probabilité commune $h(\tilde{x}; \tilde{\theta}) = f(\tilde{y}_i | \tilde{z}_i; \tilde{\theta})g(\tilde{z}_i; \tilde{\phi})$. Supposons que l'unité i soit échantillonnée avec une probabilité $w(\tilde{x}_i; \tilde{\alpha})$, qui dépend des valeurs \tilde{x}_i et, probablement, d'un paramètre vectoriel inconnu $\tilde{\alpha}$. Désignons par X_i^w les valeurs enregistrées pour l'unité i es. La densité de probabilité de X_i^w est alors

$$h^w(\tilde{x}_i; \tilde{\alpha}, \tilde{\theta}) = f(\tilde{x}_i | i\text{es}) = P[i\text{es} | X(i) = \tilde{x}_i] h(\tilde{x}_i; \tilde{\theta}) / P(i\text{es})$$

$$(3.2) \quad = w(\tilde{x}_i; \tilde{\alpha}) h(\tilde{x}_i; \tilde{\theta}) / \int w(\tilde{x}_i; \tilde{\alpha}) h(\tilde{x}_i; \tilde{\theta}) d\tilde{x}_i.$$

Les paramètres cibles dans l'inférence analytique sont le paramètre vectoriel $\tilde{\theta}$ ou des fonctions de celui-ci. Soit $s = \{1, \dots, n\}$ un échantillon de taille fixe $n < N$ prélevé avec remise de telle sorte qu'à n importe quel tirage $k = 1, \dots, n, P(j\text{es}) = w(\tilde{x}_j; \tilde{\alpha}), j = 1, \dots, N$. La densité de probabilité conjointe de $\{X_i^w, i = 1, \dots, n\}$ est alors $\prod_{i=1}^n h^w(\tilde{x}_i; \tilde{\alpha}, \tilde{\theta})$ de

où

$$\Sigma_{ij} = \text{COV}[(\tilde{Y}_i, \tilde{Y}_j)'], \quad i, j = 1, 2, \quad B_{12} = \Sigma_{12} \Sigma_{22}^{-1} \quad \text{et} \quad \tilde{B}_{12} = \Sigma_{12} \tilde{\Sigma}_{22}^{-1}.$$

En ce qui a trait à l'expression explicite de B_{12} , voir Holt, Smith et Winter (1980).

2.3 Estimateurs corrigés selon le plan (ECP)

Supposons qu'on puisse faire abstraction du processus d'échantillonnage. Désignons par $f_N(\tilde{\theta}; Y)$ la fonction de vraisemblance logarithmique de θ que l'on obtiendrait dans le cas d'un recensement. Soient $h_N(Y | Z, Y_s; \tilde{\theta}_2)$ la distribution conditionnelle de Y étant donné Z et Y_s et $E_{h_N}(\cdot | Z, Y_s)$ le terme d'espérance mathématique pour h_N . L'ECP $\tilde{\theta}^{ND}$ de $\tilde{\theta}$, tel que le propose Chambers (1986), est défini par l'équation

$$E_{h_N}[-\ell_N(\tilde{\theta}^{ND}) | Z, Y_s] = \min\{E_{h_N}[-\ell_N(\tilde{\theta}) | Z, Y_s]; \tilde{\theta} \in \Theta\}. \tag{2.5}$$

Notons que l'espérance mathématique $E^{ND}(\theta) = E_{h_N}[\ell_N(\theta) | Z, Y_s]$ dépend du paramètre vectoriel $\tilde{\theta}_1$ de la distribution conditionnelle $f(Y | Z; \tilde{\theta}_1)$. On calcule l'estimateur $\tilde{\theta}^{ND}$ de l'équation (2.5) en substituant $\tilde{\theta}_1$ à $\hat{\theta}_1$, où $\hat{\theta}_1$ est l'EMV de $\tilde{\theta}_1$ établi à partir des données (Y_s, Z) . Par des opérations algébriques simples, on peut montrer qu'en ce qui concerne le modèle normal multidimensionnel de la section 2.2, les ECP de μ_Y et Σ_Y sont identiques aux EMV définis en (2.4). Par ailleurs, un avantage probable de cette méthode est qu'elle peut être appliquée à d'autres fonctions de perte.

2.4 Pseudo-méthode de vraisemblance

Le trait marquant de cette méthode est qu'elle utilise les probabilités d'échantillonnage pour estimer les équations de vraisemblance pour le recensement. Les équations estimées sont ensuite maximisées par rapport au paramètre vectoriel étudié. Il n'est pas nécessaire de connaître la valeur des variables de plan, bien que l'étude empirique montre qu'il est possible d'accroître l'efficacité des estimateurs si on connaît la valeur de ces variables pour la population. Supposons que les valeurs \tilde{Y}_i pour la population sont tirées de façon indépendante d'une distribution commune $f(\tilde{Y}; \tilde{\theta})$ et soit $\ell_N(\tilde{\theta}; \tilde{Y}) = \sum_{i=1}^N \log f(\tilde{Y}_i; \tilde{\theta})$ la fonction de vraisemblance logarithmique pour le recensement. Suivant certaines conditions de régularité, l'EMV $\tilde{\theta}$ est une solution des équations

$$\tilde{U}(\tilde{\theta}) = d\ell_N(\tilde{\theta}; \tilde{Y})/d\tilde{\theta} = \sum_{i=1}^N \tilde{u}(\tilde{\theta}; \tilde{Y}_i) = \tilde{0}, \tag{2.6}$$

où "d" est l'opérateur de dérivée et $\tilde{u}(\tilde{\theta}; \tilde{Y}_i) = d \log f(\tilde{Y}_i; \tilde{\theta})/d\tilde{\theta}$. Le pseudo-EMV de $\tilde{\theta}$ est défini comme la solution de $\tilde{U}(\tilde{\theta}) = \tilde{0}$, où $\tilde{U}(\tilde{\theta})$ est un estimateur convergent selon le plan de $\tilde{U}(\tilde{\theta})$ en ce sens que $\text{plim}_{n \rightarrow \infty, N \rightarrow \infty} [\tilde{U}(\tilde{\theta})] = \tilde{0}$ pour tous $\tilde{\theta} \in \Theta$. L'estimateur courant de $\tilde{U}(\tilde{\theta})$ est l'estimateur de Horvitz-Thompson (Horvitz et Thompson 1952), de sorte que le pseudo-EMV de $\tilde{\theta}$ est la solution de $\tilde{U}(\tilde{\theta}) = \sum_{i=1}^N w_i^* \tilde{u}(\tilde{\theta}; \tilde{Y}_i) = \tilde{0}$, où w_i^* est égal à $[1/P(\text{ies})]$ pour l'échantillonnage sans remise et à $w_i^* = (1/nP_i)$ pour l'échantillonnage avec remise. En ce qui concerne le modèle normal multidimensionnel, les pseudo-EMV de $\tilde{\mu}_Y$ et de Σ_Y sont

$$\tilde{\mu}_Y = \sum_{i=1}^N w_i^* \tilde{Y}_i / \sum_{i=1}^N w_i^*, \quad \Sigma_Y = \sum_{i=1}^N w_i^* (\tilde{Y}_i - \tilde{\mu}_Y)(\tilde{Y}_i - \tilde{\mu}_Y)' / \sum_{i=1}^N w_i^*. \tag{2.7}$$

On obtient le pseudo-EMV de la matrice de coefficients, B_{12} , par la formule $\tilde{B}_{12} = \tilde{\Sigma}_{12} \tilde{\Sigma}_{22}^{-1}$. L'ouvrage de Skinner et coll. (1989) renferme divers exemples d'application de cette méthode dans des modèles différents. Notons aussi les ouvrages de Binder (1983), de Chambliss et Boyle (1985), de Roberts, Rao et Kumar (1987) et de Pfeffermann (1988).

On dit que le processus d'échantillonnage est "négligeable" ou "neutre" si l'inférence fondée sur (2.1) équivaut à celle fondée sur (2.2). C'est exactement ce qu'on observe pour les plans d'échantillonnage qui dépendent seulement des variables de plan, \tilde{Z} , puisque dans ce cas, $P(\tilde{I} \mid Y, Z; \tilde{g}_1) = P(\tilde{I} \mid Z; \tilde{g}_1)$. Rubin (1976), Little (1982) et Sugden et Smith (1984) définissent et illustrent les conditions exactes qui doivent exister pour que l'on puisse faire abstraction du processus d'échantillonnage.

Les difficultés que soulève l'estimation par la méthode du maximum de vraisemblance (MMV) sur la base de données d'enquêtes complexes apparaissent maintenant plus clairement compte tenu des équations (2.1) et (2.2). En premier lieu, il faut connaître toutes les variables de plan pertinentes ainsi que leur valeur pour la population. Or, comme on l'affirme souvent dans les ouvrages statistiques (voir Pfreffermann 1993 pour une bibliographie), ce n'est pas nécessairement le cas. En deuxième lieu, il faut que le processus d'échantillonnage soit "négligeable" au sens où nous l'entendons ci-dessus ou encore, que les probabilités $P(\tilde{I} \mid Y, Z; \tilde{g})$ soient modélisées et incluses dans la fonction de vraisemblance. Enfin, l'utilisation de la MMV exige la spécification de la densité de probabilité conjointe $f(Y, Z; \tilde{g}, \tilde{\varphi}) = f(Y \mid Z; \tilde{g}, \tilde{\varphi}) g(Z; \tilde{\varphi})$.

2.2 MMV exacte basée sur une factorisation de la fonction de vraisemblance

Anderson (1957) a été le premier à proposer la factorisation de la fonction de vraisemblance pour des données tirées d'une distribution normale multidimensionnelle. La factorisation est réalisable lorsque les observations sont organisées selon un plan hiérarchique, c'est-à-dire lorsqu'il est possible de constituer l'ensemble de variables d'enquête X_1, \dots, X_p de manière que X_j soit observée pour toutes les unités pour lesquelles X_{j+1} est observée, $j = 1, \dots, (p - 1)$. Rubin (1974) étend l'analyse à d'autres distributions et à des structures de données plus générales. Holt, Smith et Winter (1980) reprennent le raisonnement pour l'estimation de coefficients de régression basée sur des données d'enquêtes à plan de sondage complexe.

Supposons que le processus d'échantillonnage est "négligeable", de sorte que l'inférence peut reposer sur la distribution conjointe $f(X_s, Z; \tilde{g}, \tilde{\varphi}) = f(X_s \mid Z; \tilde{g}_1) g(Z; \tilde{\varphi})$. On peut en conséquence factoriser la fonction de vraisemblance comme suit:

$$(2.3) \quad L(\tilde{\theta}, \tilde{\varphi}; Y_s, Z) = L(\tilde{\theta}_1; X_s \mid Z) L(\tilde{\varphi}; Z).$$

En supposant que les paramètres $\tilde{\theta}_1$ et $\tilde{\varphi}$ soient distincts au sens où l'entend Rubin (1976), on peut calculer l'estimation la plus vraisemblable de chacun de ces paramètres à l'aide des composantes respectives de l'équation (2.3).

Si on applique (2.3) dans le cas où (\tilde{Y}', \tilde{Z}') sont tirées d'une distribution normale multidimensionnelle, on obtient les EMV suivants pour $\tilde{\mu}_Y = E(\tilde{Y})$ et $\Sigma_Y = V(\tilde{Y})$ (Anderson 1957).

$$(2.4) \quad \tilde{\mu}_Y = \tilde{Y}_s + \tilde{g}(\tilde{Z} - \tilde{z}_s); \quad \Sigma_Y = s_{YY} + \tilde{g}[s_{ZZ} - s_{ZZ}\tilde{g}'],$$

où $(\tilde{Y}_s, \tilde{z}_s) = \Sigma_{i=1}^n (\tilde{Y}_i, \tilde{z}_i)/n$, $\tilde{Z} = \Sigma_{i=1}^n \tilde{z}_i/N$, $s_{ZZ} = \Sigma_{i=1}^n \tilde{z}_i(\tilde{Z} - \tilde{z}_i)(\tilde{Z} - \tilde{z}_i)'/N$, $s_{ZZ} = \Sigma_{i=1}^n (\tilde{z}_i - \tilde{z}_s)(\tilde{z}_i - \tilde{z}_s)'/n$.

On déduit directement de (2.4) l'estimation la plus vraisemblable de la matrice des coefficients, B_{12} , de la régression à plusieurs variables de \tilde{Y}_1 par rapport à \tilde{Y}_2 , où $\tilde{Y}' = (\tilde{Y}'_1, \tilde{Y}'_2)$. Par conséquent, si

$$\Sigma_Y = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

distributions pondérées (Rao 1965, 1985; Patil et Rao 1978), la méthode en question utilise les probabilités d'échantillonnage. Nous l'illustrons ici en nous servant de distributions normales et de deux plans d'échantillonnage différents; l'étude montre que la méthode est efficace dans ces conditions. Un autre avantage que semble faire ressortir l'étude empirique par rapport à la méthode proposée est sa robustesse devant une mauvaise spécification des variables de plan. La section 2 sert à passer en revue les diverses méthodes d'estimation fondées sur le principe du maximum de vraisemblance qui sont traitées dans la littérature statistique. Dans la section 3, nous exposons dans ses grandes lignes la nouvelle méthode. La section 4 sert à décrire l'étude empirique et à en exposer les résultats. La section 5 renferme les conclusions.

2. EXAMEN DES MÉTHODES TRAITÉES DANS LA LITTÉRATURE STATISTIQUE

Dans cette section, nous examinons sommairement les méthodes que l'on propose dans les ouvrages statistiques pour établir des estimations à partir de données d'enquête selon le principe du maximum de vraisemblance; il peut s'agir de la méthode du maximum de vraisemblance ou de versions approchées de cette méthode. Afin de mieux saisir la difficulté du problème, nous allons d'abord étudier la notion de **plan d'échantillonnage "neutre"**. Pour une analyse plus détaillée de la méthode du maximum de vraisemblance et des autres méthodes destinées à l'inférence analytique pour les enquêtes par sondage, le lecteur se référera à Pfeffermann (1993).

2.1 Plans d'échantillonnage "neutres" et informatifs

Supposons que $\tilde{Z}' = (Z^1, \dots, Z^K)$ désigne K variables de plan (variables auxiliaires) qui servent à l'élaboration du plan de sondage et désignons par $Z = (\tilde{z}_1, \dots, \tilde{z}_N)$ la matrice $N \times K$ des valeurs de \tilde{Z} , de sorte que \tilde{z}_i est le vecteur se rapportant à l'unité i . Les variables de plan peuvent comprendre des variables indicatrices de strate ainsi que des caractères quantitatifs pour les unités et les grappes. Désignons par $\tilde{Y}' = (Y_1, \dots, Y_p)$ les variables de réponse. Pour des raisons de commodité, nous supposons que \tilde{Y} est indépendant de \tilde{Z} bien que, comme nous le mentionnons plus bas et le soulignons dans l'étude empirique, les probabilités d'échantillonnage puissent dépendre directement des valeurs de Y . La matrice des valeurs des variables de réponse, $Y = (y_1, \dots, y_N)$, peut être décomposée sous la forme $Y = [Y_s, Y_\delta]$, où $Y_s = \{\tilde{y}_i, i \in s\}$ et $Y_\delta = \{\tilde{y}_i, i \notin s\}$. Soit $\tilde{I} = (I_1, \dots, I_N)$ un vecteur d'indicateurs d'inclusion dans l'échantillon conçu de telle sorte que $I_i = 1$ lorsque $i \in s$ et $I_i = 0$ dans le cas contraire.

Comme nous l'avons vu dans l'introduction, l'inconvénient majeur de l'estimation par la méthode du maximum de vraisemblance (MMV) sur la base de données d'enquêtes complexes est qu'en règle générale, $f(Y_s; \tilde{\lambda}^*) \neq \int f(Y; \tilde{\lambda}) dY_\delta$, où le symbole $f(\cdot, \cdot)$ représente la densité de probabilité. Nous avons aussi vu dans l'introduction qu'il est parfois possible de résoudre cette difficulté en modélisant la distribution conjointe de Y et de Z . Par conséquent, supposons que l'on connaît la valeur de \tilde{Z} pour chaque unité de la population et que l'on observe \tilde{Y} uniquement pour les unités de l'échantillon. On peut donc exprimer la densité de probabilité conjointe de l'ensemble des données par la formule

$$f(Y_s, \tilde{Z}; \tilde{\theta}, \tilde{\phi}, \tilde{\psi}) = \int f(Y_s, Y_\delta | Z; \tilde{\theta}_1) P(\tilde{I} | Y, Z; \tilde{\psi}) g(Z; \tilde{\phi}) dY_\delta. \tag{2.1}$$

Si on fait abstraction du processus d'échantillonnage dans l'inférence, celle-ci reposera donc sur la distribution conjointe de Y_s et de Z ; autrement dit, on ne tient pas compte du terme de probabilité, $P(\tilde{I} | Y, Z; \tilde{\psi})$, qui figure dans le membre de droite de l'équation (2.1). L'inférence sera donc basée sur l'équation

$$f(Y_s, Z; \tilde{\theta}, \tilde{\phi}) = \int f(Y_s, Y_\delta | Z; \tilde{\theta}_1) g(Z; \tilde{\phi}) dY_\delta. \tag{2.2}$$

$$\hat{\mu}_Y = \bar{y}_s = \sum_{i=1}^n y_i/n; \hat{\sigma}_Y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n = s_Y^2 \quad (1.1)$$

comme EMV de μ_Y et de σ_Y^2 . De toute évidence, $E_M(\hat{\mu}_Y) = \mu_Y$ et $E_M\{[n/(n-1)]\hat{\sigma}_Y^2\} = \sigma_Y^2$ où $E_M\{\cdot\}$ représente l'espérance mathématique selon le modèle, les unités de l'échantillon étant invariables.

Cas B – L'échantillon est tiré avec des probabilités proportionnelles à z_i avec remise de telle sorte qu'à chaque tirage $k = 1, \dots, n$, $P_i = P(\text{ies}) = z_i / \sum_{j=1}^N z_j$. Les données connues de l'analyse sont $\{y_i, z_i, \text{ies}\}$ et $\{z_{n+1}, \dots, z_N\}$. Supposons que $\text{Corr}(Y, Z) > 0$. Cela implique que $P(Y_i > \mu_Y | \text{ies}) > 1/2$ étant donné que le plan d'échantillonnage favorise le tirage d'unités pour lesquelles la valeur de Z est donc celle de Y sont élevées. De toute évidence, les estimateurs définis en (1.1) ne peuvent plus être des EMV dans les circonstances.

La situation que nous venons de décrire correspond à l'exemple "classique" des données manquantes qui fait souvent l'objet d'analyses dans les ouvrages statistiques (Anderson 1957). Les EMV de μ_Y et de σ_Y^2 sont, dans ce cas-ci,

$$\hat{\mu}_Y = \bar{y}_s + b(\bar{Z} - \bar{z}_s); \hat{\sigma}_Y^2 = s_Y^2 + b^2(S_Z^2 - s_Z^2), \quad (1.2)$$

où $\bar{Z} = \sum_{i=1}^N z_i/N$, $\bar{z}_s = \sum_{i=1}^n z_i/n$, $b = \sum_{i=1}^n (y_i - \bar{y}_s)(z_i - \bar{z}_s) / \sum_{i=1}^n (z_i - \bar{z}_s)^2$, $S_Z^2 = \sum_{i=1}^N (z_i - \bar{Z})^2/N$ et $s_Z^2 = \sum_{i=1}^n (z_i - \bar{z}_s)^2/n$. Notons que dans ce cas-ci, il est possible d'atténuer les effets d'échantillonnage en modélisant la distribution conjointe de la variable de réponse Y et de la variable de plan Z . On dit alors du processus d'échantillonnage qu'il est "neutre" (voir section 2.1).

Cas C – Mêmes conditions que dans le cas B, sauf que seules les valeurs d'échantillon $\{(y_i, z_i, \text{ies})\}$ et les probabilités d'échantillonnage $\{P_i, \text{ies}\}$ sont connues. Bien que les valeurs de z_i , $i = 1, \dots, N$, soient connues au moment de l'échantillonnage, il arrive souvent que les fichiers destinés à l'analyse secondaire ne contiennent pas d'information sur les variables de plan ou les probabilités de sélection relatives aux unités qui ne font pas partie de l'échantillon.

Les estimateurs définis en (1.2) ne sont plus pertinents dans les circonstances puisque la moyenne et la variance de Z pour la population sont inconnues. Toutefois, en ce qui concerne les grandes populations, pour lesquelles $Z \approx$ constante, on peut définir un EMV approximatif de μ_Y par l'expression $\mu_Y^* = \bar{y}_s + b^*(1/N - \bar{P}_s)$, où $\bar{P}_s = \sum_{i=1}^n P_i/n$ et $b^* = \sum_{i=1}^n (y_i - \bar{y}_s)(P_i - \bar{P}_s) / \sum_{i=1}^n (P_i - \bar{P}_s)^2$. L'estimateur μ_Y^* se justifie par le fait que $P_i = Z_i/N\bar{Z}$, de telle sorte que pour $\bar{Z} =$ constante, (Y_i, P_i) est un vecteur de valeurs tirées d'une distribution normale bidimensionnelle où $\bar{P} = \sum_{i=1}^n P_i/N = 1/N$. Cet estimateur est une illustration du cas où, comme le recommande Rubin (1985), on utilise les probabilités d'échantillonnage en remplacement des variables de plan lorsque l'information sur celles-ci est incomplète.

Il est possible d'obtenir un EMV approximatif pour le cas C en appliquant ce que les auteurs appellent la pseudo-méthode de vraisemblance. Celle-ci est décrite dans la section 2; disons qu'elle consiste essentiellement à maximiser un estimateur convergent selon le plan de la fonction de caractérisation du recensement, c'est-à-dire de la fonction de caractérisation que l'on aurait obtenue dans le cas d'un recensement. Cette fonction n'est aucunement influencée par le plan. La pseudo-méthode de vraisemblance produit pour le cas C les estimateurs suivants:

$$\hat{\mu}_Y = \bar{y}^{ps} = \sum_{i=1}^n w_i^* y_i / \sum_{i=1}^n w_i^*; \hat{\sigma}_Y^2 = s_Y^2 = \sum_{i=1}^n w_i^* (y_i - \bar{y}^{ps})^2 / \sum_{i=1}^n w_i^*, \quad (1.3)$$

où $w_i^* = (1/nP_i)$. Comme \bar{y}^{ps} et s_Y^2 sont des estimateurs convergents selon le plan de $Y = \sum_{i=1}^N y_i/N$ et de $S_Y^2 = \sum_{i=1}^N (y_i - Y)^2/N$ respectivement, ils sont aussi des estimateurs convergents de μ_Y et de σ_Y^2 en ce sens que $\text{plim}_{n \rightarrow \infty} (\bar{y}^{ps}, s_Y^2) = (\mu_Y, \sigma_Y^2)$. Dans cet article, nous étudions une autre méthode d'estimation fondée sur le principe du maximum de vraisemblance; cette méthode est, en théorie, applicable même lorsque l'analyste dispose pour seule information des données de l'échantillon. Inspirée de la théorie des

Estimation par la méthode du maximum
de vraisemblance dans des enquêtes
par sondage complexes

ABBA M. KRIEGER et DANNY PFEFFERMANN¹

RÉSUMÉ

L'estimation de vraisemblance maximale fondée sur des données d'échantillon complexe nécessite une plus grande modélisation en raison de l'information contenue dans le plan d'échantillonnage. Par ailleurs, on peut appliquer des pseudo-méthodes du maximum de vraisemblance, qui consistent à maximiser des valeurs estimées de la fonction de caractérisation du recensement. Dans cet article, nous examinons quelques-unes des méthodes décrites dans les ouvrages spécialisés et nous les comparons à une nouvelle méthode qui tire son origine de la notion de "distributions pondérées". Les comparaisons portent principalement sur des situations où la totalité ou une fraction des variables de plan sont inconnues ou mal spécifiées. Les résultats obtenus avec la nouvelle méthode sont encourageants mais précisons que notre analyse se limite actuellement à des situations simples.

MOTS CLÉS: Estimateurs corrigés selon le plan; plans "neutres" et plans informatifs; pseudo-méthode de vraisemblance; distributions pondérées.

1. INTRODUCTION

On se sert souvent des données d'enquête pour faire de l'inférence analytique sur des paramètres de modèle comme les moyennes, les coefficients de régression, les probabilités par case, etc. Les modèles se rapportent aux données de la population; c'est pourquoi on les appelle modèles de recensement. Lorsqu'elles sont appliquées à des données d'enquête, les méthodes du maximum de vraisemblance "classiques" présentent un inconvénient en ceci que le modèle qui est valable pour l'échantillon peut être très différent de celui qui est valable pour la population à cause des effets d'échantillonnage.

Pour illustrer le problème et quelques-unes des solutions proposées dans la littérature statistique, prenons un exemple simple. Une population U est constituée de N unités désignées par $\{1, \dots, N\}$. À chaque unité i correspond un vecteur (Y_i, Z_i) de mesures indépendantes tirées d'une distribution normale bidimensionnelle de moyenne $\mu' = (\mu_Y, \mu_Z)$ et de matrice des variances-covariances

$$\Sigma = \begin{bmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{bmatrix}.$$

Les valeurs (y_i, z_i) sont observées pour un échantillon s de $n < N$ unités prélevées suivant un plan d'échantillonnage probabiliste. On cherche à estimer μ_Y et σ_Y^2 . Nous considérons trois cas qui se distinguent l'un de l'autre par le processus d'échantillonnage et le genre de données disponibles.

Cas A – L'échantillon est tiré suivant un plan d'échantillonnage aléatoire simple avec remise et seules les valeurs $\{y_i, z_i, i \in s\}$ sont connues. En désignant les unités de l'échantillon par $\{1, \dots, n\}$, nous pouvons dire que $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} N(\mu_Y, \sigma_Y^2)$ ce qui donne

¹ Abba M. Krieger, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104. Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905.

BIBLIOGRAPHIE

- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GODAMBE, V.P., et THOMPSON, M.E. (1984). Robust estimation through estimating equations. *Biometrika*, 71, 115-125.
- GODAMBE, V.P., et THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- LIANG, K.-Y., et WACHTLAWI, M.A. (1990). Extension of the Stein Estimating Procedure through the use of estimating functions. *Journal of the American Statistical Association*, 85, 435-440.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., et SINGH, M.P. (Eds.) (1987). *Small Area Statistics An International Symposium*. New York: Wiley.
- SÄRNDAL, C.-E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. *Bulletin de l'Institut International de Statistique*, 49, 494-513.
- SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning Inference*, 7, 155-170.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, A.C., MANTEL, H.J., et THOMAS, B.W. (1991). Time series generalizations of Fay-Herriot estimation for small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

produits à l'aide de cette méthode avec un ensemble d'intervalles basés uniquement sur un modèle en supposant la normalité des erreurs et en utilisant une variable r . Si les deux ensembles d'intervalles sont très différents, on aura lieu de douter de la validité des intervalles basés à la fois sur un modèle et sur un plan; cependant, d'autres recherches seront nécessaires avant que nous puissions résoudre cette question de façon satisfaisante.

Binder (1983) propose une autre méthode d'estimation de la variance par rapport au sujet qui nous occupe. On peut en effet estimer la variance de plan de h^* comme estimateur de g^* à γ^N au moyen des techniques courantes basées sur un plan, en substituant γ_s à γ^N , et on peut ensuite calculer la variance de γ_s comme estimateur de γ^N au moyen d'une linéarisation de Taylor de h^* par rapport à γ^N . On pourrait recourir aussi à la linéarisation de Taylor pour calculer l'estimateur de la variance d'une fonction de γ_s qui sert d'estimateur pour la fonction équivalente de γ^N .

4. SUJETS DE RECHERCHE POUR L'AVENIR

Nous venons de voir comment la méthode décrite dans cet article peut servir à estimer des moyennes de populations finies ou, plus généralement, des fonctions de paramètres de régression linéaire. Il est normal de se demander si cette méthode peut aussi servir à l'estimation d'autres types de paramètres de population finie, comme les fonctions de distribution et les quantiles, ou à l'estimation pour petits domaines et si cela nécessite des ajustements et lesquels.

Prenons le cas particulier de l'estimation d'une fonction de distribution à un point précis. Il y a deux façons d'intégrer de l'information de covariables dans un modèle. La première est de poser explicitement la probabilité comme une fonction des covariables, comme dans le modèle logistique. La seconde, qui est courante dans l'estimation de fonctions de distribution, comme dans Chambers et Dunstan (1986), Rao, Kovar et Mantel (1990) et d'autres, est de poser que les résidus d'une régression de la variable observée sur les covariables sont indépendants et identiquement distribués suivant une loi indéterminée. Selon cette méthode, le paramètre étudié doit être une fonction du paramètre de population finie. Peut-on se servir de cette méthode pour l'estimation de fonctions de distribution ou de quantiles?

Une autre question majeure dans le domaine des sondages est l'estimation pour petits domaines, c'est-à-dire l'estimation de totaux, de moyennes ou de proportions pour des sous-ensembles de la population finie. Platek, Rao, Särndal et Singh (1987) présentent une analyse détaillée de la question. Une façon simple d'adapter la méthode décrite dans la section 1 à l'estimation pour petits domaines serait de l'appliquer séparément dans chacun des domaines étudiés; il s'agirait alors, en quelque sorte, d'une estimation par régression généralisée avec stratification *a posteriori*. Notons que dans les circonstances, il faut connaître les totaux des covariables pour chaque domaine étudié. Une pratique courante dans l'estimation pour petits domaines consiste à "emprunter" de l'information dans les domaines voisins par l'intermédiaire d'un modèle qui met en relation les petits domaines et certaines covariables de même que les petits domaines entre eux. Singh, Mantel et Thomas (1991) font une bonne analyse de la question. Une méthode qui s'est avérée très utile est l'estimation empirique de Bayes fondée sur des modèles à effets aléatoires, que l'on doit à Fay et Herriot (1979). Liang et Wacziarg (1990) examinent des fonctions d'estimation destinées à des modèles empiriques de Bayes. Est-il possible de construire des modèles qui visent à "emprunter" de l'information dans les domaines voisins de manière que les paramètres étudiés deviennent des fonctions d'un paramètre de population?

REMERCIEMENTS

L'auteur remercie l'arbitre et le rédacteur en chef pour leurs observations utiles.

3. ESTIMATION DE LA VARIANCE ET INTERVALLES DE CONFIANCE

Une méthode de production d'intervalles de confiance qui serait conforme à l'idée générale des fonctions d'estimation serait de construire un pivot normal multidimensionnel asymptotique basé sur h^* ainsi qu'un estimateur de la variance correspondant. Les régions de confiance approximatives pour $\hat{\gamma}_N$ correspondraient alors à des régions de probabilité de la distribution normale multidimensionnelle estimée de ce pivot approximatif. Or, nous ne nous intéressons pas à $\hat{\gamma}_N$ mais à une fonction non injective de cet estimateur. Nous allons opter pour une méthode plus simple, à savoir l'estimation directe de la variance de $\hat{\gamma}_{\text{AREG}}$. Särndal, Swensson et Wretman (1989) se sont penchés sur l'estimation de la variance pour $\hat{\gamma}_{\text{GREG}}$, défini en (5), lorsque le second terme de l'expression est nul. Nous avons vu dans la section 2 que l'estimateur $\hat{\gamma}_{\text{AREG}}$ a précisément cette forme. L'estimateur de la variance de Särndal, Swensson et Wretman peut s'écrire

$$\hat{V}_g = \sum_{i \in S} \sum_{j \in S} \Delta_{ij} g_{is} \hat{e}_{is} g_{js} \hat{e}_{js}, \tag{8}$$

où $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij}$, π_{ij} , étant la probabilité de plan que les individus i et j fassent partie de l'échantillon s , g_{is} est le i -ième élément du vecteur ligne $w_T^s (W_T^s V_s^{-1} \Pi_s^{-1} W_s^s)^{-1} W_s^s V_s^{-1}$, et $\hat{e}_{is} = (y_i - x_i^T \hat{\gamma}_s) / \pi_i$. Voir Särndal, Swensson et Wretman (1989) pour une analyse détaillée des propriétés de modèle et de plan de \hat{V}_g en (8). Notons que $\hat{\gamma}_{\text{AREG}}$ en (7) peut s'écrire $\hat{\gamma}_{\text{AREG}} = \sum_{i \in S} g_{is} y_i / \pi_i$ et

$$\hat{\gamma}_{\text{AREG}} - \hat{\gamma}_N = \sum_{i \in S} g_{is} \hat{e}_{iN} = w_T^N (\hat{\gamma}_s - \hat{\gamma}_N),$$

où $\hat{e}_{iN} = (y_i - w_i^T \hat{\gamma}_N) / \pi_i$. Or, si $V_{iN}^N = W^N a$, nous avons $w_T^N = 1^T V_N V_N^{-1} W_N^N / N = a^T W_T^N V_N^{-1} W_N^N / N$, de sorte que pour de grands échantillons, g_{is} tendra vers $1/N$ pour $i \in S$. La variance de plan de $\hat{\gamma}_{\text{AREG}}$ est donc approximativement égale à

$$\sum_{j \in P} \sum_{i \in P} \Delta_{ij} \hat{e}_{iN} \hat{e}_{jN} / N^2,$$

où $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)$, et cette expression peut être estimée au moyen de l'équation

$$\hat{V}_1 = \sum_{i \in S} \sum_{j \in S} \Delta_{ij} \hat{e}_{is} \hat{e}_{js} / N^2. \tag{9}$$

On parle de l'estimateur ci-dessus dans des ouvrages pionniers sur l'estimateur par régression généralisé; voir, par exemple, Särndal (1981, 1982). Nous pouvons considérer l'estimateur \hat{V}_g défini en (8) comme une version de \hat{V}_1 corrigée en fonction des valeurs observées de g_{is} , $i \in S$. Särndal, Swensson et Wretman (1989) montrent que \hat{V}_g est souvent non biaisé ou quasi non biaisé selon le modèle pour l'incertitude quadratique moyenne de modèle de $\hat{\gamma}_{\text{AREG}}$ tout en étant convergent selon le plan pour la variance de plan du même estimateur. On peut maintenant construire des intervalles de confiance approximatifs pour $\hat{\gamma}_N$ en se fondant sur une approximation normale de la distribution de $(\hat{\gamma}_{\text{AREG}} - \hat{\gamma}_N) / \{ \hat{V}_g \}^{1/2}$. Cette méthode a un fondement asymptotique tant du point de vue du plan que du point de vue du modèle, et il est nécessaire de vérifier si elle est bien adaptée à certains échantillons de taille finie. À cette fin, on peut, par exemple, comparer un ensemble d'intervalles de confiance

2. COMPARAISON AVEC L'ESTIMATEUR PAR RÉGRESSION GÉNÉRALISÉ

Soit W_N la matrice de plan pour le modèle élargi, c'est-à-dire

(6)
$$W_N = (V^N \mathbf{1}_N, X^N).$$

Pour les besoins de l'analyse, nous supposons que $V^N \mathbf{1}_N$ n'est pas inclus dans l'espace vectoriel engendré par les colonnes de X^N . De même, posons W_s comme la version augmentée de X_s et $\gamma_s, \hat{\gamma}_s$ et $\hat{\gamma}_s$ comme les versions augmentées de $\beta_s, \hat{\beta}_N$, et $\hat{\beta}_s$ respectivement. Pour des raisons de commodité, nous désignerons notre estimateur de la moyenne de population comme l'estimateur par régression augmenté,

(7)
$$\hat{y}^{\text{AREG}} = W_N^T \hat{\gamma}_s.$$

Nous allons montrer tout d'abord que \hat{y}^{AREG} est aussi une sorte d'estimateur de différence généralisé. D'après (6), si u est un vecteur de grandeur appropriée dont le premier élément est égal à un et les autres nuls, alors $W_N u = V^N \mathbf{1}_N$ et $W_s u = V_s \mathbf{1}_s$. Par conséquent,

$$\mathbf{1}_s^T \Pi_s^{-1} W_s^T \hat{\gamma}_s = u^T W_s^T V_s^{-1} \Pi_s^{-1} W_s^T \hat{\gamma}_s = u^T W_s^T V_s^{-1} \Pi_s^{-1} y_s = \mathbf{1}_s^T \Pi_s^{-1} y_s$$

et il s'ensuit que le second terme de l'estimateur par régression généralisé de l'équation (5), où $\hat{\beta}_s$ est remplacé par $\hat{\gamma}_s$, est égal à 0. En deuxième lieu, comparons \hat{y}^{AREG} (équ. 7) et \hat{y}^{GREG} (équ. 5). De longs calculs nous amènent à l'équation suivante:

$$\hat{y}^{\text{AREG}} = \hat{x}_N \hat{\beta}_s + (c_1/c_2) \mathbf{1}_s^T \Pi_s^{-1} (y_s - X_s \hat{\beta}_s)/N,$$

où

$$c_1 = \mathbf{1}_N^T (V^N \mathbf{1}_N - X^N (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s)$$

et

$$c_2 = \mathbf{1}_s^T \Pi_s^{-1} (V_s \mathbf{1}_s - X_s (X_s^T V_s^{-1} \Pi_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} \mathbf{1}_s).$$

Exprimé de cette manière, \hat{y}^{AREG} ressemble étroitement à \hat{y}^{GREG} , sauf en ce qui concerne le facteur de pondération placé devant le second terme. Il ne semble pas possible d'expliquer de façon heuristique le poids (c_1/c_2). Cependant, on constate que c_1 est précisément la somme des résidus d'une régression pondérée des v_i sur les x_i fondée sur l'échantillon s et que c_2 ressemble à un estimateur d'Horvitz-Thompson de c_1 , sauf que les résidus dépendent aussi de l'échantillon s . Dans le cas de grands échantillons tirés de grandes populations, (c_1/c_2) devrait tendre vers 1. En comparant \hat{y}^{AREG} avec \hat{y}^{GREG} , nous pouvons dire que le premier est plutôt fondé sur un plan tandis que le second est plutôt fondé sur un modèle. Il est vrai que \hat{y}^{AREG} est convergent selon le plan, mais \hat{y}^{GREG} a en plus une caractéristique de plan pour échantillon de taille finie en ceci que $\hat{\gamma}_s$ est la solution d'une équation d'estimation qui est non biaisée selon le plan pour l'équation à définition de paramètre de $\hat{\beta}_N$. Les équations de Thompson (1984, 1986), traitées dans Godambe et Thompson

où $y_i^T = (y_1, \dots, y_N, V_N)$, V_N est une matrice diagonale ayant pour éléments v_1, \dots, v_N , et X_N est une matrice à N lignes, la i -ième ligne étant x_i^T .

Or, β_N est inconnu. Godambe et Thompson (1986) ont défini et élaboré des méthodes d'estimation optimale simultanée de β et de β_N à partir du modèle et du plan de sondage. Désignons les données d'une enquête par sondage par $X_s = \{(i, y_i), i \in s\}$.

Pour l'estimation simultanée de β et de β_N , nous considérons des fonctions d'estimation $h(X_s, \beta)$ telles que $E_p(h) = g^*$ en (1), où E_p désigne l'espérance par rapport au plan de sondage. Une fonction h^* de cette classe est appelée optimale si, pour toutes les autres fonctions h de la même classe, $E_p(h) - E_p(h^*) \{h^* h^T\} = E_p(h^* h^*^T) - E_p(h^* h^*^T)$ est définie non négative. Le théorème 1 de Godambe et Thompson (1986) montre que la fonction optimale h^* est définie par l'expression

$$h^*(X_s, \beta) = \sum_{i \in s} (y_i - x_i^T \beta) x_i / \pi_i v_i, \quad (3)$$

où π_i est la probabilité, selon le plan de sondage, que l'individu i fasse partie de l'échantillon s . Nous désignerons la racine de cette fonction par $\hat{\beta}_s$, c'est-à-dire,

$$\hat{\beta}_s = (X_s^T \Pi_s^{-1} V_s^{-1} X_s)^{-1} X_s^T \Pi_s^{-1} V_s^{-1} y_s, \quad (4)$$

où y_s est le vecteur des y_i s pour $i \in s$, Π_s et V_s sont des matrices diagonales ayant pour éléments π_i et v_i respectivement, $i \in s$, et X_s est une matrice ayant pour lignes x_i^T , $i \in s$.

Jusqu'à maintenant, il n'a été question que de l'estimation de β ou β_N . En réalité, nous cherchons à estimer \hat{y}_N , la moyenne de population des y_i s. Pour cela, nous pouvons utiliser, par exemple, un estimateur par régression généralisé,

$$y_{\text{REG}} = x_N^T \hat{\beta}_s + \mathbf{1}_s^T \Pi_s^{-1} (y_s - X_s \hat{\beta}_s) / N, \quad (5)$$

où $\mathbf{1}_s$ est un vecteur de uns dont la grandeur équivaut à la taille de l'échantillon s . Särndal, Swensson et Wretman (1992), notamment, analysent cet estimateur. Le premier terme de l'estimateur offre de bonnes propriétés par rapport au modèle tandis que le second terme offre de bonnes propriétés par rapport au plan. Toutefois, l'adéquation de y_{REG} au modèle et au plan ne dépend pas de la forme particulière de $\hat{\beta}_s$, et rien, à première vue, ne nous empêche de croire que l'on pourrait remplacer $\hat{\beta}_s$ dans (5) par un estimateur de β basé uniquement sur un modèle. L'optimalité selon le plan de $\hat{\beta}_s$ est apparemment peu pertinente.

L'estimateur que nous proposons ici établit une relation plus étroite entre le modèle hypo-thétique et le paramètre de population finie y_N . Comme β_N en (2) est estimé de manière optimale par $\hat{\beta}_s$ en (4), les fonctions de $\hat{\beta}_s$ sont estimées de manière optimale par la fonction équivalente de $\hat{\beta}_s$. Si $y_N = u^T \hat{\beta}_N$ pour un vecteur u quelconque, nous estimerons y_N au moyen de $u^T \hat{\beta}_s$. Un tel vecteur existe si et seulement si $V_N \mathbf{1}_N$ est inclus dans l'espace vectoriel engendré par les colonnes de X_N , auquel cas, si $V_N \mathbf{1}_N = X_N u$, nous pouvons avoir $u = X_N^T V_N^{-1} X_N u / N = x_N$. En revanche, si $V_N \mathbf{1}_N$ n'est pas dans l'espace vectoriel engendré par les colonnes de X_N , nous allons l'y inclure. Ce faisant, nous sacrifions quelque peu l'efficacité du modèle, bien que la version élargie demeure acceptable par rapport au modèle initial. Nous insistons moins sur l'efficacité du modèle afin d'accroître le rapport avec des populations finies. Il est intéressant de noter que lorsque les variances du modèle ne dépendent pas de i , notre méthode aboutit à l'inclusion d'une constante arbitraire dans le modèle de régression.

La méthode proposée ici se rapproche beaucoup de celle de Little (1983), qui suggère de recourir à l'estimation basée sur des modèles, et seulement les modèles qui produisent des estimateurs asymptotiquement convergents selon le plan. Par ailleurs, Isaki et Fuller (1982) proposent de se limiter aux plans pour lesquels l'estimateur fondé sur un modèle est asymptotiquement convergent selon le plan.

Estimation pour population finie à l'aide de fonctions d'estimation HAROLD J. MANTEL¹ RÉSUMÉ

Godambe et Thompson (1986) définissent et élaborent des méthodes d'estimation optimale simultanée de paramètres de superpopulation et de population finie à partir d'un modèle de superpopulation et d'un plan de sondage. Dans leurs travaux, ils définissent le paramètre de population finie, θ_N , comme la solution de l'équation d'estimation optimale pour le paramètre de superpopulation θ ; cependant, un autre paramètre de population finie, ϕ , peut être tout aussi intéressant. Nous proposons d'élargir le modèle de superpopulation de telle manière que le paramètre ϕ soit une fonction connue de θ_N , c.-à-d. $\phi = f(\theta_N)$. Alors, ϕ est estimé de manière optimale par $f(\theta_s)$, où, d'après Godambe et Thompson (1986), θ_s est l'estimateur optimal de θ_N étant donné l'échantillon s et le plan de sondage.

MOTS CLÉS: Fonctions d'estimation; estimateur linéaire généralisé; paramètre de population finie.

1. ESTIMATION D'UNE MOYENNE

Dans cet article, nous traitons de l'estimation d'un paramètre de population finie tel que la moyenne établie à partir d'une enquête par sondage. Il est aussi question d'un modèle de régression hypothétique pour superpopulation qui met en relation la variable étudiée et des covariables connues. Le but de cet exercice est d'obtenir une méthode d'estimation qui offre de bonnes propriétés aussi bien par rapport au plan de sondage que par rapport au modèle hypothétique. Cette étude s'inspire des travaux de Godambe et Thompson (1986).

Nous supposons que nous avons une population finie d'individus $P = \{i: i = 1, \dots, N\}$. À chaque individu i sont associées une variable inconnue y_i et un vecteur de covariables, x_i . Celui-ci peut être connu pour tous $i \in P$ ou seulement pour les i faisant partie de l'échantillon tandis que la moyenne de la population, x_N , est connue. Si nous posons E_m comme l'espérance par rapport au modèle de superpopulation, les hypothèses du modèle sont les suivantes:

- (i) y_i et y_j sont indépendantes pour $i \neq j$
- (ii) $E_m(y_i) = x_i^T \beta$ pour un vecteur réel inconnu β
- (iii) $E_m(y_i - x_i^T \beta)^2 = \sigma^2 v_i$, $i = 1, \dots, N$, pour des valeurs v_i connues et une valeur σ^2 inconnue.

D'après Godambe et Thompson (1986), nous définissons un paramètre de population finie β_N comme la solution de l'équation d'estimation linéairement optimale

$$(1) \quad g^* = \sum_N^{i=1} (y_i - x_i^T \beta) x_i / v_i = 0,$$

c'est-à-dire,

$$(2) \quad \hat{\beta}_N = (X_N^T V_N^{-1} X_N)^{-1} X_N^T V_N^{-1} y_N,$$

¹ H. J. Mantel, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

- RAO, J.N.K., et WU, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data. *Bulletin de l'Institut Internationale de Statistique*.
- SHAO, J. (1991). *L*-statistics in complex survey problems. Rapport technique, Université d'Ottawa, Ottawa.
- SHAO, J., et WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.
- SHAO, J., et WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics*, 20 (à paraître).
- SITTER, R.R. (1992). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, (à paraître).
- WANG, J.C., et WU, C.F.J. (1991). An approach to the construction of asymmetrical orthogonal arrays. *Journal of the American Statistical Association*, 86, 450-456.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other positional measures. *Journal of the American Statistical Association*, 47, 635-646.
- WU, C.F.J. (1989). Construction of 2^{m4^n} designs via a grouping scheme. *Annals of Statistics*, 17, 1880-1885.
- WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

Le tableau 2 indique les valeurs résultant de la simulation pour le biais relatif, le coefficient de variation, les taux d'erreur des extrêmes inférieure (I) et supérieure (S), ainsi que les étendues standardisées. En premier lieu, nous notons que l'estimateur de la variance selon la méthode de d'auto-amorçage (3.5) affiche un biais relatif plus grand et un coefficient de variation (CV) légèrement plus élevé que sa variante obtenue en remplaçant θ par θ^* ; biais relatif de 12,6% comparativement à 7,5%, et CV de 52% comparativement à 48% pour $m_h = n_h - 1 = 4$. En revanche, l'estimateur de la variance BRR affiche le biais relatif le plus faible (3,1%) et le plus petit CV (3,1%), tandis que l'estimateur de la variance basé sur l'intervalle de Woodruff comporte un biais relatif plus faible (4,2%) et un CV comparable (47%). En deuxième lieu, les taux d'erreur des extrêmes inférieure et supérieure sont voisins du niveau nominal (5%) pour les intervalles selon les méthodes d'auto-amorçage et BRR, tandis que les taux d'erreur sont légèrement inégaux pour l'intervalle de Woodruff ($I = 4,2\%$ et $S = 5,6\%$). Enfin, nous observons que les étendues standardisées sont à peu près égales pour toutes les méthodes. Dans l'ensemble, l'estimateur de la variance selon la méthode d'auto-amorçage et les intervalles d'auto-amorçage basés sur la méthode des centiles ne se sont pas révélés meilleurs que l'estimateur de la variance Woodruff et l'intervalle de Woodruff.

REMERCIEMENTS

Les travaux de J.N.K. Rao ont bénéficié d'une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

DEY, A. (1985). *Orthogonal Fractional Factorial Designs*. New Delhi: Wiley Eastern.

EFFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

GUPTA, V.K., et NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

GURNEY, M., et JEWETT, R.S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, 70, 819-821.

HANSEN, M., et TEPPING, B.J. (1985). Estimation for variance in NAEP. Note de service non-publiée, Westat, Washington, D.C.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La Revue Canadienne de Statistique*, 16, 25-45.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

MCCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Revue de l'Institut Internationale de Statistique*, 37, 239-264.

RAO, J.N.K. (1988). Variance estimation in sample surveys. Dans *Handbook of Statistics*, Vol. 6, (Éds. P.R. Krishnaiah et C.R. Rao). Amsterdam: Elsevier Science, 427-447.

RAO, J.N.K., et WU, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., et WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

(1 - α) pour θ est alors donné par $\{\theta - t_{S_j}^*(\theta), \theta - t_{S_j}^*(\theta)\}$, où $t_{S_j}^*$ et $t_{S_j}^*$ sont les points $\alpha/2$ inférieur et supérieur de l'histogramme de la méthode d'auto-amorçage de t^* . Cet histogramme d'auto-amorçage peut aussi permettre l'établissement d'intervalle de confiance unilatéraux. Des travaux empiriques de Kovar, Rao et Wu (1988) ayant porté sur des fonctions lisses indiquent que l'intervalle de confiance qu'on en déduit avec $m_h = n_h - 1$ permet la détermination des taux d'erreur tant à l'extrémité inférieure qu'à l'extrémité supérieure de manière plus efficace qu'avec l'intervalle jackknife $\{\theta - z_{\alpha/2} s_j(\theta), \theta + z_{\alpha/2} s_j(\theta)\}$, mais que le taux d'erreur total ne se différencie pas de celui obtenu avec cette dernière méthode, c.-à-d. que pour des intervalles bilatéraux, les deux méthodes ont une performance semblable pour ce qui est de la probabilité d'inclure la valeur vraie. Si une transformation stabilisatrice de la variance peut être trouvée, par exemple la transformation \tanh^{-1} appliquée au coefficient de corrélation estimé, le problème des taux d'erreur inégaux aux deux extrémités posé par l'intervalle jackknife semble se corriger. On peut donc conclure que l'intervalle jackknife, ou tout autre intervalle basé sur la théorie normale, s'il se fonde sur de telles transformations, peut être utile quand les transformations sont connues, tandis que la méthode d'auto-amorçage offre une solution de rechange quand de telles transformations n'existent pas ou sont inconnues.

Nous présentons maintenant les résultats d'une étude de simulation limitée relative à la performance de la méthode d'auto-amorçage proposée dans le cas de la médiane. En utilisant la population de base 1 de Hansen-Topping avec $L = 32$ strates (voir Kovar et coll. 1988, sections 3 et 6 pour avoir des détails), nous avons produit 500 échantillons aléatoires simples stratifiés indépendants avec $n_h = 5$, puis nous avons calculé le biais relatif et le coefficient de variation (stabilité relative) de l'estimateur de la variance basé sur l'intervalle de Woodruff avec $\alpha = 0.1$ (voir Kovar et coll. 1988, éq. (2.8)), de l'estimateur de la variance BRR (3.3), ainsi que de l'estimateur de la variance selon la méthode d'auto-amorçage (3.5) et de sa variante obtenue en remplaçant θ par $\theta^{(.)}$. Nous avons utilisé $m_h = n_h - 1$ et $n_h - 3$, ainsi que $B = 500$ échantillons répétés selon la méthode d'auto-amorçage pour chaque échantillon, tandis que les échantillons répétés BRR ont été obtenus à partir d'un tableau orthogonal, au moyen de 250 exécutions. Nous avons fait une approximation de l'erreur quadratique moyenne vraie de θ en choisissant 10,000 échantillons aléatoires stratifiés indépendants. Nous avons aussi calculé les taux d'erreur à chaque extrémité (taux nominal de 5% à chaque extrémité) et les étendues standardisées de l'intervalle de confiance basé sur la loi normale, en utilisant l'estimateur de la variance BRR, l'intervalle de Woodruff et l'intervalle selon la méthode d'auto-amorçage obtenu d'après la méthode des centiles au moyen de l'histogramme de $\theta^{(1)}, \dots, \theta^{(\theta^*)}$ pour chaque échantillon.

Tableau 2

Biais relatif en % et CV en % d'estimateurs de la variance, et taux d'erreur et étendues standardisées d'intervalles de confiance (niveau nominal de 5% à chaque extrémité), pour la médiane, selon un échantillonnage aléatoire simple stratifié $L = 32, n_h = 5$

Méthode	Biais relatif (%)	CV %	Taux d'erreur	Étendue standardisée
			I	S
Woodruff	4.2	47	4.2	5.6
BRR	3.1	31	5.0	5.0
Auto-amorçage*:				
$m_h = 4$	12.6	52	5.0	5.2
$m_h = 2$	13.0	54	5.0	4.8

* Les résultats relatifs à la variance de l'estimateur de la variance selon la méthode d'auto amorçage sont indiqués entre parenthèses.

être facilement construits pour n'importe quelle combinaison, comme on peut le voir dans Wu (1989). Autrement dit, le nombre de répétitions peut être aussi petit que possible. Par conséquent, il est possible de constituer un vaste ensemble de tableaux orthogonaux mixtes en vue d'une utilisation pratique si n_h se limite à 2, 3 ou 4.

La méthode BRR et les prolongements examinés jusqu'ici n'exigent le prélèvement que d'une seule unité (upé) par strate pour chaque échantillon répété. Pour le cas où n_h est grand, disons supérieur à 3, Sitter (1992) a proposé l'emploi de multi-tableaux orthogonaux, de façon à permettre que le nombre d'unités rééchantillonnées par strate soit supérieur à un. Cela pourrait exiger moins d'échantillons répétés et pourrait englober des cas où il n'existe pas de tableaux orthogonaux de force deux, par exemple si $n_h = 6$.

3.3 Méthode d'auto-amorçage

La méthode d'auto-amorçage, dans le cas des variables iid, a été abondamment étudiée (Efron 1982). Rao et Wu (1987) ont proposé une extension englobant les plans d'échantillonnage à plusieurs degrés stratifiés, mais limitée aux statistiques lisses $\theta = g(X)$. Cette extension exigeait, pour que l'estimation de la variance soit valide dans le cas où n_h est petit, qu'une correction d'échelle soit effectuée, semblable à celle décrite à la section 3.2. Toutefois, les auteurs ont omis de tenir compte du fait que la correction d'échelle devrait porter sur les poids de l'enquête w_{hik} plutôt que sur les valeurs y_{hik} directement, comme ils le proposaient. Par conséquent, leur méthode ne peut être étendue au quantile $\theta = F^{-1}(p)$. Nous présentons maintenant une méthode générale englobant à la fois les statistiques lisses et les statistiques non lisses, pour des tailles arbitraires n_h . En voici le fonctionnement: (i) Sélectionner un échantillon aléatoire simple de m_h grappes avec remise parmi les n_h grappes de l'échantillon, indépendamment pour chaque h . Soit m_{hi}^* le nombre de fois que la (hi) -ième grappe de l'échantillon est choisie ($\sum_i m_{hi}^* = m_h$). Définissons ainsi les poids propres à la méthode d'auto-amorçage:

$$w_{hik}^* = \left[\{ 1 - (m_h/(n_h - 1))^{1/2} + (m_h/(n_h - 1))^{1/2} (n_h/m_h) m_{hi}^* \} w_{hik} \right] \quad (3.4)$$

Si la (hi) -ième grappe n'est pas incluse dans l'échantillon d'auto-amorçage, $m_{hi}^* = 0$ et le second terme de (3.4) disparaît. Si m_h est choisi de manière à être inférieur ou égal à $n_h - 1$, les poids d'auto-amorçage w_{hik}^* sont tous positifs si $w_{hik} (0$ pour tout (hik) es. Calculer θ^* , l'estimateur de θ , selon la méthode d'auto-amorçage, en utilisant les poids w_{hik}^* dans la formule de θ . La médiane d'auto-amorçage, par exemple, est calculée comme auparavant au moyen des poids d'auto-amorçage normalisés $w_{hik}^* / \sum_s w_{hik}^*$, pourvu que tous les w_{hik}^* soient plus grands que zéro. (ii) Répéter de façon indépendante l'étape (i) un grand nombre de fois, disons B , et calculer les estimations correspondantes $\theta_{(1)}^*, \dots, \theta_{(B)}^*$.

L'estimateur d'auto-amorçage de la variance selon la méthode d'auto-amorçage $s_{\text{BOOT}}^2(\theta) = E^*(\theta^* - E^*\theta^*)^2$, est donné approximativement par

$$s_{\text{BOOT}}^2(\theta) = \frac{1}{B} \sum_{b=1}^B [\theta_{(b)}^* - \theta]^2. \quad (3.5)$$

On obtient une variante de (3.5) en remplaçant θ par $\theta_{(b)}^*$, $= \sum_b \theta_{(b)}^* / B$. Dans le cas linéaire, $s_{\text{BOOT}}^2(\theta)$ se réduit à l'estimateur "approprié" de la variance $s^2(X)$.

Rao et Wu (1987) ont obtenu des intervalles de confiance selon la méthode d'auto-amorçage pour des fonctions lisses, $\theta = g(X)$, au moyen d'une approximation de la distribution de $t = (\theta - \theta)/s_j(\theta)$ à l'aide de sa contrepartie d'auto-amorçage $t^* = (\theta^* - \theta)/s_j(\theta^*)$, où $s_j^2(\theta^*)$ est obtenue de (3.2) en remplaçant w_{hik} par w_{hik}^* . Un intervalle de confiance bilatéral au niveau

$\Sigma_r \delta_h' \delta_{h'}' = 0$ pour tout $h \neq h'$, c.-à-d. que les colonnes de la matrice sont orthogonales. Un ensemble minimum de R demi-échantillons compensés peut être construit à partir de matrices de Hadamard ($L + 1 \leq R \leq L + 4$) en choisissant n importe quel ensemble de L colonnes excluant la colonnes contenant les valeurs $+1$.

Soit $\theta^{(r)}$ l'estimateur de θ obtenu à partir du r -ième demi-échantillon. Notons que $\theta^{(r)}$ est obtenu de θ en remplaçant le poids du (h/k) -ième élément par $2w_{hik}$ ou 0 selon que la (hi) -ième grappe est choisie ou non dans le demi-échantillon. Un estimateur de la variance selon la méthode BRR pour θ est donné par

$$s_{BRR}^2(\theta) = \frac{1}{R} \sum_{r=1}^R (\theta^{(r)} - \theta)^2. \quad (3.3)$$

Plusieurs variantes de $s_{BRR}^2(\theta)$ sont aussi disponibles; par exemple, θ peut être remplacé par $\theta(\cdot) = \Sigma_r \theta^{(r)}/R$. Dans le cas linéaire, $\theta = X$, tous les estimateurs de la variance selon la méthode BRR sont réduits à l'estimateur "approprié" de la variance, $s^2(X)$, comme pour la méthode du jackknife.

Krewski et Rao (1981) ont établi la convergence de $s_j^2(\theta)$ et de $s_{BRR}^2(\theta)$ pour des statistiques lisses $\theta = g(X)$, à mesure que L augmente. Rao et Wu (1985) ont effectué une analyse du second ordre et ont montré que $s_{BRR}^2(\theta)$ et $s_j^2(\theta)$ ne sont pas asymptotiquement équivalents à des termes du second ordre, contrairement à $s_j^2(\theta)$ et $s_j^2(\theta)$. Shao et Wu (1992) ont établi la convergence de $s_{BRR}^2(\theta)$ pour les quantiles, $\theta = F^{-1}(p)$.

La méthode BRR a été étendue au cas de $n_h = p > 2$ grappes par strate, p étant premier ou une puissance de nombre premier (Gurney et Jewett 1975), mais le nombre de répétitions, R , nécessaires est beaucoup plus grand que dans le cas de $n_h = 2$. Dans beaucoup de plans d'enquête, les n_h ne sont pas égaux. Pour tenir compte de ce cas général où les n_h sont inégaux, Gupta et Nigam (1987), ainsi que Wu (1991), ont proposé l'utilisation de tableaux orthogonaux mixtes de force deux pour le tirage d'échantillons répétés compensés, où n_h est le nombre de symboles dans la h -ième colonne du tableau. L'orthogonalité du tableau assure que les échantillons répétés sont compensés. Comparativement au cas où les n_h sont égaux, la correction des poids de l'enquête est plus complexe. Une méthode appropriée a été énoncée par Wu (1991). D'après sa formule (6), deux corrections distinctes devraient être appliquées aux unités sélectionnées et non sélectionnées pour faire partie de chaque échantillon répété. Un traitement algébrique simple de l'équation (6) de Wu montre que w_{hik} devient $w_{hik}' = [1 + (n_h - 1)^{1/2}] w_{hik}$ ou $w_{hik}' = [1 - (n_h - 1)^{1/2}] w_{hik}$ selon que le (hik) -ième élément est choisi ou non pour faire partie de l'échantillon répété. (Notons que $w_{hik}' = 2$ et $w_{hik}' = 0$ pour $n_h = 2$). Le calcul qui reste pour obtenir $\theta^{(r)}$ et $s_{BRR}^2(\theta)$ est le même qu'en (3.3). En outre, ces poids de l'enquête modifiés peuvent être appliqués à $\theta = F^{-1}(p)$, et au cas plus général $\theta = T(F)$, où T est une fonctionnelle de F . Il suffit de remplacer w_{hik} dans (2.4) par w_{hik}' ou w_{hik}'' selon que le (hik) -ième élément est inclus ou non dans le r -ième échantillon répété pour obtenir $F^{(r)}$ de F pour le r -ième échantillon répété, ainsi que $\theta^{(r)} = T(F^{(r)})$. Le calcul de l'estimateur de la variance selon la méthode BRR est le même qu'en (3.3).

L'utilisation de tableaux orthogonaux mixtes pose deux problèmes. Premièrement, le tableau peut être de taille élevée pour un n_h général. Deuxièmement, il n'existe pas de tableaux orthogonaux pour toutes les combinaisons de n_h . Une solution pratique consiste à réunir les n_h unités primaires d'échantillonnage (upé) de la strate h en deux ou quatre groupes d'upé, puis d'appliquer la méthode BRR aux groupes en traitant ces derniers comme des unités. Cette extension est appelée la méthode BRR à unités groupées. Comme l'a montré Wu (1991), la perte d'efficacité peut être relativement faible, par rapport à la méthode BRR complète, si les groupements sont faits judicieusement. Par exemple, il faut d'avantage de groupes si n_h est grand et si les unités formant la strate sont plus hétérogènes. Pour $n_h = 2, 3$ ou 4, de nombreux tableaux orthogonaux mixtes ont été construits (voir, par exemple, Dey 1985 et Wang et Wu 1991). Si n_h ne peut prendre que les valeurs 2 ou 4, des tableaux orthogonaux saturés peuvent

Tableau 1

Biais et biais relatif en % (entre parenthèses) de l'estimateur de la variance jackknife pour la médiane selon un échantillonnage par grappes stratifié ($n_h = 2, L = 32$) et certaines valeurs de corrélation intra-grappes identiques, M et certaines tailles de grappes identiques, M

ρ	M				
	1	10	20	30	50
0	7.5(116)	.28(41)	.09(29)	.04(15)	.01(15)
0.05	-	.22(27)	.09(18)	.05(12)	.03 (8)
0.10	-	.28(28)	.10(14)	.06 (9)	.02 (3)
0.20	-	.31(22)	.11(10)	.08 (8)	.03 (3)
0.30	-	.32(18)	.11 (7)	.07 (5)	.01 (1)
0.50	-	.44(17)	.15 (6)	.11 (5)	.04 (2)

Dans l'étude de simulation, nous avons produit des échantillons de grappes stratifiées $\{y_{hik}, k = 1, \dots, M, i = 1, \dots, n_h; h = 1, \dots, L\}$ en utilisant le modèle d'erreurs emboîtées $y_{hik} = \mu_h + a_{hi} + e_{hik}$ avec $a_{hi} \sim N(0, \sigma_{ah}^2)$ et $e_{hik} \sim N(0, \sigma_{eh}^2)$, où la taille des grappes, M , est supposée identique pour toutes les grappes (hi), et où les corrélations intra-grappes, $\sigma_{ah}^2 / (\sigma_{ah}^2 + \sigma_{eh}^2) = \rho_h$, sont supposées égales pour toutes les strates h (c.-à-d. (i.e., $\rho_h = \rho$). Les poids de l'enquête normalisés sont donnés par w_{hik} , avec $w_{hik} = w_h / (n_h M)$, où w_h désigne la taille relative de la strate h . Le nombre de strates $L (= 32)$, les moyennes des strates, μ_h , les variances $\sigma_h^2 = \sigma_{ah}^2 + \sigma_{eh}^2$ et les tailles w_h ont été choisis de manière à correspondre à des populations réelles observées dans une étude américaine intitulée "US National Assessment of Educational Progress Study" (Hansen et Tepping 1985). Nous avons produit 1,000 échantillons par grappes stratifiées indépendants avec $n_h = 2$ pour chaque combinaison sélectionnée (ρ, M), puis nous avons calculé le biais et le biais relatif de l'estimateur de la variance selon la méthode du jackknife, $s_j^2(\theta)$, pour la médiane: biais de $[s_j^2(\theta)] = \sum_i s_{ji}^2(\theta) / 1,000 - \text{MSE}(\theta)$, où $s_{ji}^2(\theta)$ est la valeur de $s_j^2(\theta)$ pour le i -ième échantillon simulé ($i = 1, \dots, 1,000$); biais relatif de $[s_j^2(\theta)] = \text{biases de } [s_j^2(\theta)] / \text{MSE}(\theta)$. Nous avons calculé $\text{MSE}(\theta)$ à partir d'un ensemble indépendant de 10,000 échantillons par grappes stratifiées pour chaque (ρ, M) : $\text{MSE}(\theta) = \sum_i (\theta_i - \theta)^2 / 10,000$, où θ_i est la valeur de θ pour le i -ième échantillon simulé, $\theta = \sum \theta_i / t$ et $t = 1, \dots, 10,000$.

Le tableau 1 indique les valeurs simulées du biais et du biais relatif (entre parenthèses) de l'estimateur de la variance selon la méthode du jackknife pour certaines combinaisons de ρ et de M . En premier lieu, nous notons que dans le cas spécial de l'échantillonnage aléatoire simple stratifié ($\rho = 0, M = 1$), le biais relatif est très élevé (116%), ce qui confirme la non-convergence de $s_j^2(\theta)$ dans ce cas. En deuxième lieu, nous observons que le biais et le biais relatif diminuent tous deux quand M augmente, pour un ρ donné. De plus, pour une taille de grappes donnée M , le biais augmente généralement avec ρ , mais le biais relatif, en fait, décroît, parce que $\text{MSE}(\theta)$ augmente plus vite que le biais quand ρ s'accroît. Il est réconfortant de constater, en effet, que le biais relatif ne dépasse pas 10% pour $M \geq 30$ et $\rho \geq 0.10$ ou $M \geq 20$ et $\rho \geq 0.20$.

3.2 Méthode BRR

La méthode BRR ("balanced repeated replication") a été proposée par McCarthy (1969) pour le cas spécial important de $n_h = 2$ grappes par strate. Un ensemble de R demi-échantillons compensés (répétitions) est formé en retranchant une grappe de l'échantillon dans chaque strate. Cet ensemble peut être défini par une matrice de plan $R \times L$ dénotée (δ_h^r) , $1 \leq r \leq R, 1 \leq h \leq L$ avec $\delta_h^r = +1$ ou -1 selon que la première ou la deuxième grappe de l'échantillon appartenant à la h -ième strate se trouve dans le r -ième demi-échantillon, et

3.1 Jackknife

Pour la simplicité, supposons que $\theta = g(Y)$ soit une fonction lisse du total estimé Y . Soit $\theta^{(g)} = g(Y^{(g)})$ l'estimateur de θ obtenu de l'échantillon après avoir omis les données provenant de la j -ième grappe sélectionnée dans la g -ième strate ($j = 1, \dots, n_g; g = 1, \dots, L$), où

$$Y^{(g)} = \sum_{h \neq g} w_{hik} Y_{hik} + \sum_{\substack{(gik) \in s \\ i \neq j}} \left\{ \frac{n_g}{n_g - 1} w_{gik} \right\} Y_{gik}. \quad (3.1)$$

Notons que $Y^{(g)}$ est obtenu en remplaçant le poids du (gik) -ième élément par $n_g w_{gik}/(n_g - 1)$, mais en conservant les poids originaux, w_{hik} , pour $h \neq g$. L'expression suivante donne un estimateur courant de la variance de θ selon la méthode du jackknife avec suppression d'une grappe:

$$s_j^2(\theta) = \sum_{L=1}^g \frac{n_g}{n_g - 1} \sum_{f=1}^j (\theta^{(gf)} - \theta)^2. \quad (3.2)$$

Deux variantes de $s_j^2(\theta)$ sont obtenues en remplaçant θ dans (3.2) par $\theta^{(g\cdot)} = \sum_j \theta^{(gj)}/n_g$ et $\theta^{(\cdot\cdot)} = \sum_g \sum_j \theta^{(gj)}/n$, où $n = \sum_g n_g$. Dans le cas linéaire, $\theta = Y$, tous les estimateurs de la variance selon la méthode du jackknife sont réduits à l'estimateur "approprié" de la variance, $s^2(Y)$, donné par (2.3). Rao et Wu (1987) ont effectué une analyse de second ordre des estimateurs de la variance de rééchantillonnage quand θ est exprimé sous forme d'une fonction lisse de totaux, Y . Leurs principales conclusions concernant la méthode du jackknife sont les suivantes: 1) Différents estimateurs de la variance selon la méthode du jackknife sont toujours égaux à des termes d'ordre supérieur à mesure que s accroît le nombre de strates, L . 2) Dans le cas important où $n_h = 2$ pour tous les h , l'estimateur de la variance selon la méthode de linéarisation, $s_L^2(\theta)$, ainsi que l'estimateur de la variance selon la méthode du jackknife, sont asymptotiquement égaux à des termes d'ordre supérieur à mesure que s accroît le nombre de strates, L . Le choix entre les deux méthodes devrait dépendre davantage de facteurs opérationnels que de critères statistiques.

La méthode du jackknife courante avec une suppression, dans le cas d'observations indépendantes et identiquement distribuées (iid), comporte un inconvénient, à savoir que contrairement à la méthode d'auto-amorçage, elle n'offre pas d'estimateur convergent de la variance pour des statistiques non lisses, comme la médiane. Shao et Wu (1989), ont montré que cette déficience peut être corrigée par l'emploi d'une forme plus générale de la méthode du jackknife, appelée méthode du jackknife avec d suppressions, dans laquelle le nombre d'observations supprimées, d , dépend d'une mesure du degré de lissage de la statistique. En particulier, pour les quantiles de l'échantillon, la méthode du jackknife avec d suppressions, où d satisfait à la condition $n^{1/2}/d \rightarrow 0$ et $n - d \rightarrow \infty$ quand $n \rightarrow \infty$ produit des estimateurs convergents de la variance dans le cas d'observations iid. Ce résultat donne à penser qu'un effet semblable pourrait demeurer valable dans la méthode du jackknife avec suppression d'une grappe pour l'échantillonnage à plusieurs degrés stratifié, puisque tous les éléments sélectionnés provenant d'une grappe (g) incluse dans l'échantillon sont supprimés quand on calcule $s_j^2(\theta)$ comme en (3.2). Nous nous penchons actuellement sur l'étude théorique de ce problème, mais nous avons effectué une étude de simulation limitée qui laisse croire que l'estimateur de la variance selon la méthode du jackknife avec suppression d'une grappe $s_j^2(\theta)$ pourrait se révéler un très bon estimateur. Nous allons maintenant exposer les résultats de l'étude de simulation pour la médiane, $\theta = F^{-1}(1/2)$.

on note que les r_{hi} sont des variables aléatoires indépendantes et identiquement distribuées (iid) ayant la même moyenne, Y_h , et la même variance dans chaque strate h , selon un échantillonnage avec remise des grappes. Il s'ensuit que l'expression suivante donne un estimateur sans biais de la variance de Y :

$$s^2(Y) = \sum_h^h s_{rh}^2/n_h, \quad (2.3)$$

avec

$$(n_h - 1)s_{rh}^2 = \sum_{hi=1}^n (r_{hi} - \bar{r}_h)^2.$$

Si l'échantillonnage des grappes est fait sans remise, $s^2(Y)$ produira une surestimation de la

variance vraie de Y .
Il est souvent intéressant, par ailleurs, d'estimer la fonction de distribution de la population $F(t)$, et le p -ième quantile, $\theta = F^{-1}(p)$, $0 < p < 1$, en particulier la médiane de la population $\theta = F^{-1}(1/2)$. L'estimateur de l'enquête de $F(t)$ est donné par:

$$F(t) = \sum_{(hik) \in s} w_{hik} a_{hik}, \quad (2.4)$$

où les $w_{hik} = w_{hik} / \sum_s w_{hik}$ sont les poids normalisés $\sum_s w_{hik} = 1$ et où $a_{hik} = 1$ si $Y_{hik} \leq t$, et $a_{hik} = 0$, et dans le cas contraire. Le p -ième quantile de l'échantillon est ainsi obtenu:

$$\hat{\theta} = F^{-1}(p). \quad (2.5)$$

En pratique, pour calculer $\hat{\theta}$, on classe d'abord les valeurs sélectionnées Y_{hik} en ordre croissant, disons $\{Y^{(hik)}\}$, puis on cumule les poids connexes w_{hik} jusqu'à ce qu'on croise p . Le premier $Y^{(hik)}$ que l'on rencontre après avoir croisé p est considéré comme le p -ième quantile de l'échantillon, $\hat{\theta}$. Woodruff (1952) a obtenu des intervalles de confiance pour un quantile, et Rao et Wu (1987) ont obtenu un estimateur simple de la variance en utilisant l'intervalle de Woodruff (voir aussi Kovar, Rao et Wu 1988, Francisco et Fuller 1991). Shao (1991) a examiné des statistiques L générales, notamment le courbe de Lorenz de l'échantillon et l'indice de concentration de Gini, qui sont des exemples de statistiques L lisses, et les quantiles de l'échantillon, qui sont des exemples de statistiques L non lisses.
Plusieurs paramètres non linéaires dignes d'intérêt, comme des moyennes de la population, des ratios, des coefficients de régression et de corrélation, peuvent être exprimés sous forme de fonctions lisses, $\theta = g(X)$, d'un vecteur de totaux, $X = (X_1, \dots, X_q)'$, de variables convenablement définies. On a alors un estimateur de θ sous la forme $\hat{\theta} = g(\hat{X})$. La méthode de linéarisation peut être utilisée pour estimer la variance de $g(\hat{X})$, en vertu de tout plan complexe (voir Binder 1983 et Rao 1988).

3. MÉTHODES DE RÉÉCHANTILLONNAGE

Les méthodes de rééchantillonnage, comme la méthode du jackknife et la méthode d'auto-amorçage, sont largement utilisées dans le cas de variables iid. Des modifications ou des prolongements appropriés de ces méthodes ont également été élaborés afin de permettre le traitement des données d'enquête provenant d'échantillons stratifiés ou en grappes. Nous présentons maintenant un bref compte rendu de certains travaux récents ayant porté sur trois de ces méthodes (jackknife, méthodes BRR et d'auto-amorçage), dans le contexte de l'échantillonnage à plusieurs degrés stratifié.

Le présent document fait état de certains travaux récents relatifs aux méthodes de rééchantillonnage applicables aux enquêtes complexes. Il énonce aussi quelques résultats empiriques concernant l'estimation de la variance par la méthode du jackknife et la méthode d'auto-amorçage, pour des statistiques non lissées comme la médiane, en vertu d'un échantillonnage par grappes stratifié et d'un échantillonnage aléatoire simple stratifié.

2. ÉCHANTILLONNAGE À PLUSIEURS DEGRÉS STRATIFIÉS

Les enquêtes à grande échelle utilisent souvent des plans d'échantillonnage à plusieurs degrés stratifiés comportant un grand nombre de strates, L , et un nombre relativement faible d'unités primaires d'échantillonnage (grappes), $n_h (\geq 2)$, prélevées dans chaque strate h . En fait, il est très fréquent de sélectionner $n_h = 2$ grappes dans chaque strate, pour permettre un niveau maximum de stratification des grappes, compatible avec l'obtention d'un estimateur valide de la variance. Nous supposons que le sous-échantillonnage appliqué aux grappes sélectionnées est effectué de manière à permettre une estimation sans biais des totaux des grappes Y_{hi} , $i = 1, \dots, n_h$; $h = 1, \dots, L$.

Soit $wh_{ik} (> 0)$ le poids de l'enquête associé au k -ième élément de l'échantillon (unité finale) appartenant à la i -ième grappe sélectionnée dans la h -ième strate. Souvent, les poids de base wh_{ik} sont soumis, au moment de la stratification à posteriori, à une correction visant à les rendre compatibles avec des totaux connus de variables propres à la stratification à *posteriori*. L'enquête sur la population active canadienne, par exemple, utilise un estimateur de régression généralisée pour assurer cette compatibilité. Nous omettrons toutefois cette complication dans la présente communication. L'expression suivante définit un estimateur du total de la population Y :

$$(2.1) \qquad Y = \sum_{(hik) \in s} wh_{ik} y_{hik},$$

où s dénote l'échantillon d'éléments et y_{hik} est la valeur d'un attribut d'intérêt, y , associé à l'élément $(hik) \in s$. Nous supposons un niveau de réponse complet pour chaque question. Il est de pratique courante de sélectionner les grappes avec probabilités proportionnelles à la taille (ppt) et sans remise, afin d'accroître l'efficacité des estimateurs par rapport à l'échantillonnage avec ppt avec remise, et pour éviter la possibilité d'inclure plus d'une fois la même grappe dans l'échantillon. Toutefois, à l'étape de l'estimation de la variance, les calculs sont énormément simplifiés si l'on traite l'échantillon comme si les grappes étaient sélectionnées avec remise et que le sous-échantillonnage était effectué indépendamment chaque fois qu'une grappe est choisie. Cette simplification se traduit pas une surestimation de la variance de Y , mais le biais relatif sera vraisemblablement faible si la fraction de sondage du premier degré est peu élevée dans chaque strate.

Si l'on écrit Y sous la forme

$$(2.2) \qquad Y = \sum_L y_h,$$

$$r_{hi} = \sum_k (n_h wh_{ik}) y_{hik}, \quad \bar{r}_h = \sum_i r_{hi} / n_h,$$

avec

Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes

J.N.K. RAO, C.F.J. WU et K. YUE¹

RÉSUMÉ

Les méthodes de rééchantillonnage permettant d'obtenir, par inférence, des résultats à partir de données d'enquêtes complexes incluent la méthode du jackknife, la méthode BRR ("balanced repeated replication") et la méthode d'auto-amorçage. La présente communication passe en revue certains travaux récents relatifs à ces méthodes, pour l'estimation des erreurs-types et des intervalles de confiance. Certains résultats empiriques relatifs à des statistiques non lisses sont également présentés.

MOTS CLÉS: BRR; auto-amorçage (bootstrap); jackknife; plans à plusieurs degrés stratifiés; estimation de la variance.

1. INTRODUCTION

La théorie classique de l'échantillonnage est largement consacrée à l'estimation de l'erreur quadratique moyenne (MSE) d'estimateurs sans biais ou approximativement sans biais \bar{y} d'un total Y pour l'ensemble de la population. Un estimateur de l'erreur quadratique moyenne, ou de la variance, fournit une mesure du degré d'incertitude de l'estimateur \bar{y} . Il est de pratique courante de supposer que l'estimateur \bar{y} suit approximativement une distribution normale, puis d'utiliser un intervalle de confiance bilatéral $\bar{y} \pm z_{\alpha/2} s(\bar{y})$, ou encore un intervalle de confiance unilatéral $(\bar{y} - z_{\alpha} s(\bar{y}), \infty)$ ou $(-\infty, \bar{y} + z_{\alpha} s(\bar{y}))$, où $s(\bar{y})$ est l'erreur-type de \bar{y} (c.-à-d. la racine carrée de l'estimation de l'erreur quadratique moyenne) et z_{α} est le point alpha supérieur d'une variable $N(0, 1)$. Ces intervalles comprennent le total vrai Y avec une probabilité d'environ $1 - \alpha$ si la taille de l'échantillon est grande, mais la probabilité réelle d'inclure le total vrai pourrait être sensiblement inférieure à $1 - \alpha$ si l'échantillon est petit ou si l'on compte beaucoup d'unités de mêmes grappes. Pour des statistiques non linéaires, comme des ratios ou des coefficients de régression ou de corrélation, la méthode bien connue de linéarisation (ou développement de Taylor) est souvent utilisée (voir Rao 1988 pour des applications détaillées). Des méthodes de rééchantillonnage, comme le jackknife, la méthode BRR et la méthode d'auto-amorçage, sont également employées. En fait, plusieurs organismes américains et canadiens ont adopté la méthode du jackknife pour l'estimation de la variance dans le cas d'échantillonnages à plusieurs degrés stratifiés. La méthode de linéarisation a l'avantage de pouvoir s'appliquer à des plans de sondage généraux, mais elle nécessite l'établissement d'une formule d'erreur-type distincte, $s(\bar{y})$, pour chaque statistique non linéaire \bar{y} . En revanche, les méthodes de rééchantillonnage utilisent une formule d'erreur-type unique pour toutes les statistiques \bar{y} . Toutefois, la méthode du jackknife et la méthode BRR ne sont applicables, strictement, qu'aux plans à plusieurs degrés stratifiés dans lesquels les grappes situées à l'intérieur des strates sont sélectionnées avec remise, ou dans lesquels la fraction de sondage du premier degré est négligeable. La méthode d'auto-amorçage de Rao et Wu (1987) convient à des plans plus généraux, mais elle exige d'abondants calculs et son adaptation aux plans complexes n'a pas encore été entièrement étudiée.

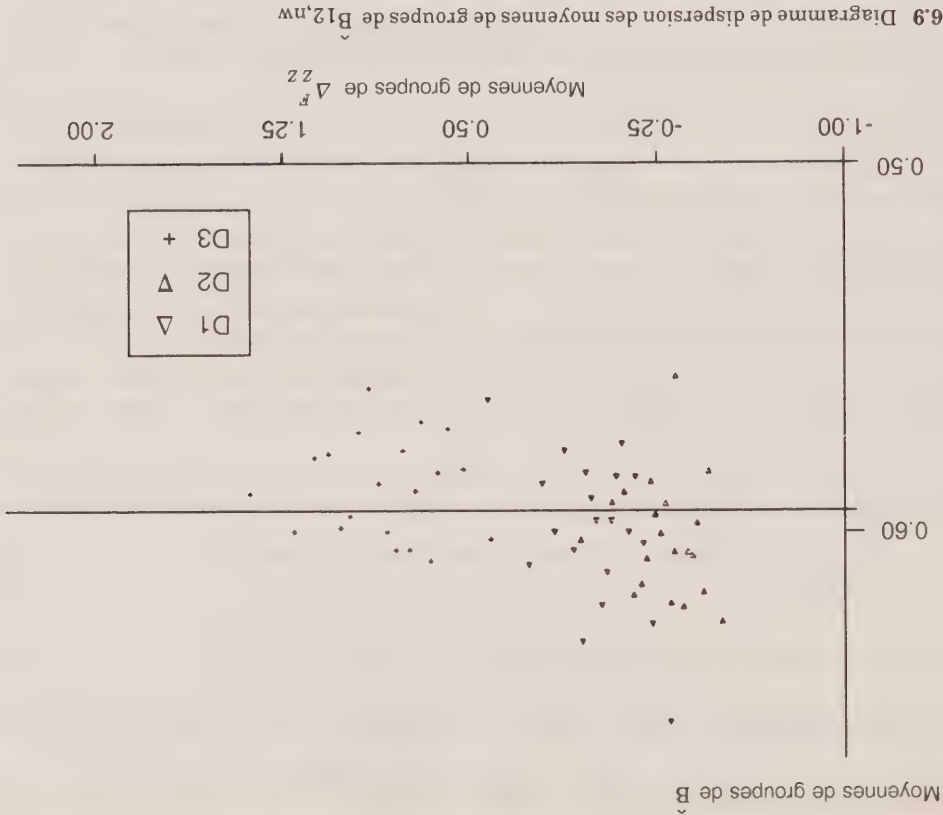
¹ J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa (Ontario) K1S 5B6, C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (Ontario) N2L 3G1, Kim Yue, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario) K1A 0T6.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric curve fitting. *Journal of the Royal Statistical Society, B*, 47, 1-52.

SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā, A*, 359-372.

- GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1211.
- GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.
- GODAMBE, V.P., et THOMPSON, M.E. (1977). Robust near optimal estimation in survey practice. *Bulletin de l'Institut Internationale de Statistique*, 47, 129-146.
- GODAMBE, V.P., et THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- HANSEN, M.H., MADOW, W.G., et TEPPING, B.J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- HOLT, D., SMITH, T.M.F., et WINTERS, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Sér. A*, 143, 474-487.
- HOLMES, D. (1987). The effect of selection on the robustness of multivariate methods. Thèse de doctorat non publiée, University of Southampton, U.K.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTZ, S., et JOHNSON, N.L. (1988). *Encyclopedia of Statistical Sciences*, (Vol. 8). New York: John Wiley, 157.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability Application*, 9, 141-142.
- NATHAN, G., et HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, B*, 42, 377-386.
- NJENGA, E.G. (1990). Robust estimation of the regression coefficients in complex surveys. Thèse de doctorat non publiée, University of Southampton.
- PARZEN, E. (1962). On the estimation of the probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.
- PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions Royal Society of London, A*, 200, 1-66.
- PFERFERNANN, D.J., et HOLMES, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, A*, 148, 268-278.
- ROYAL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYAL, R.M., et HERSON, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- ROYAL, R.M., et HERSON, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association*, 68, 890-893.
- SÄRNÄDAL, C.-E. (1980). On π -inverse weighting versus best linear weighting in probability sampling. *Biometrika*, 67, 639-650.
- SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā, C*, 39, 1-9.
- SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.



REMERCIEMENT

Les auteurs désirent remercier un arbitre anonyme de plusieurs observations utiles qui ont rehaussé la présentation de cet article. Le British Council a soutenu E. Njenga à l'aide d'une subvention.

BIBLIOGRAPHIE

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

BREWER, K.R.W. (1979). A class of robust designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W., et SÄRNDAL, C.-E. (1983). Six approaches to enumerative survey sampling. *Incomplete Data in Sample Surveys*, (Vol. 3). New York: Academic Press, 363-368.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GASSER, T., et MÜLLER, H.G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation*, (Eds. T. Gasser et M. Rosenblatt). New York: Springer-Verlag, 23-68.

GASSER, T., et ENGEL, J. (1990). The choice of weights in kernel regression estimation. *Biometrika*, 77, 377-381.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, B*, 17, 269-278.

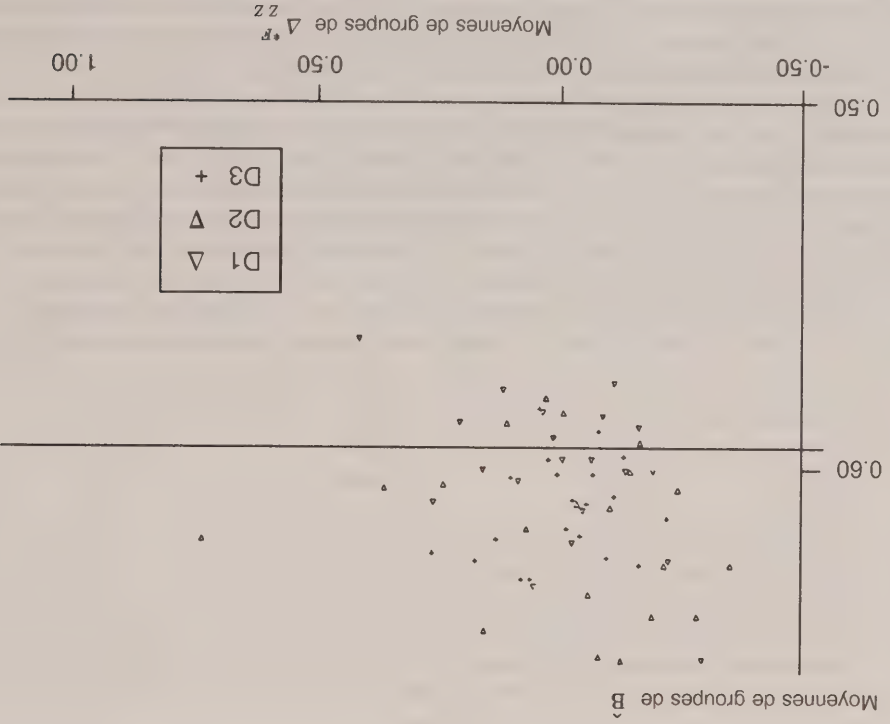
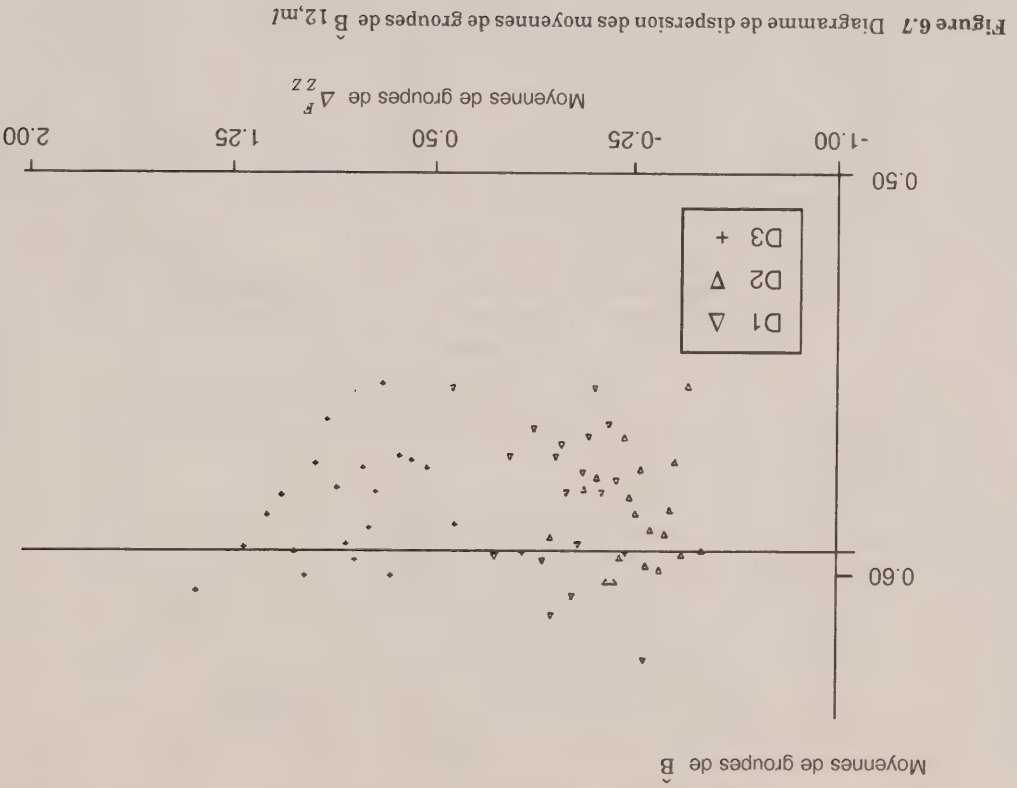


Tableau 6.9
Biais absolus inconditionnels des trois estimateurs de B_{12}
 $N = 6,962, n = 100$ Valeur vraie de $B_{12} = 0.595$

Biais absolus de			
Plan de sondage			
$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$	
D1	0.0245	0.0245	0.0056
D2	0.0260	0.0408	0.0060
D3	0.0128	0.0355	0.0072

Tableau 6.10
Écart type inconditionnel des trois estimateurs de B_{12}

Écart types			
Plan de sondage			
$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$	
D1	0.111	0.111	0.111
D2	0.106	0.132	0.108
D3	0.111	0.122	0.111

Tableau 6.11

Erreur quadratique moyenne inconditionnelle des trois estimateurs de B_{12}

Erreur quadratique moyenne			
Plan de sondage			
$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$	
D1	0.0130	0.0130	0.0121
D2	0.0120	0.0192	0.0117
D3	0.0125	0.0161	0.0123

biaisé conditionnellement dans l'un ou l'autre des plans. Enfin, nous voyons d'après la figure 6.9 que l'estimateur noyau "nw" est approximativement non biaisé conditionnellement pour les trois plans.

Ces études de simulation nous permettent de conclure que le nouvel estimateur $\hat{B}_{12,nw}$ est satisfaisant. Lorsqu'on s'écarte des hypothèses de linéarité et d'homoscédasticité, cet estimateur paraît robuste pour divers plans de sondage et semble être relativement efficace et avoir des propriétés conditionnelles et inconditionnelles raisonnables. Nous savons d'après des études antérieures que $\hat{B}_{12,pwml}$ est aussi efficace que les estimateurs pondérés en p plus classiques sur le plan inconditionnel et qu'il a des propriétés conditionnelles de beaucoup supérieures. Compte tenu de ce que le nouvel estimateur $\hat{B}_{12,nw}$ semble avoir, selon cette étude, de meilleures propriétés que l'estimateur "pwml", qui avait été choisi pour représenter le groupe d'estimateurs pondérés en p à cause de son rendement dans d'autres études de simulation, il est permis de croire que $\hat{B}_{12,nw}$ pourrait être utilisé dans des études analytiques où on s'intéresse à un petit nombre de paramètres clés.

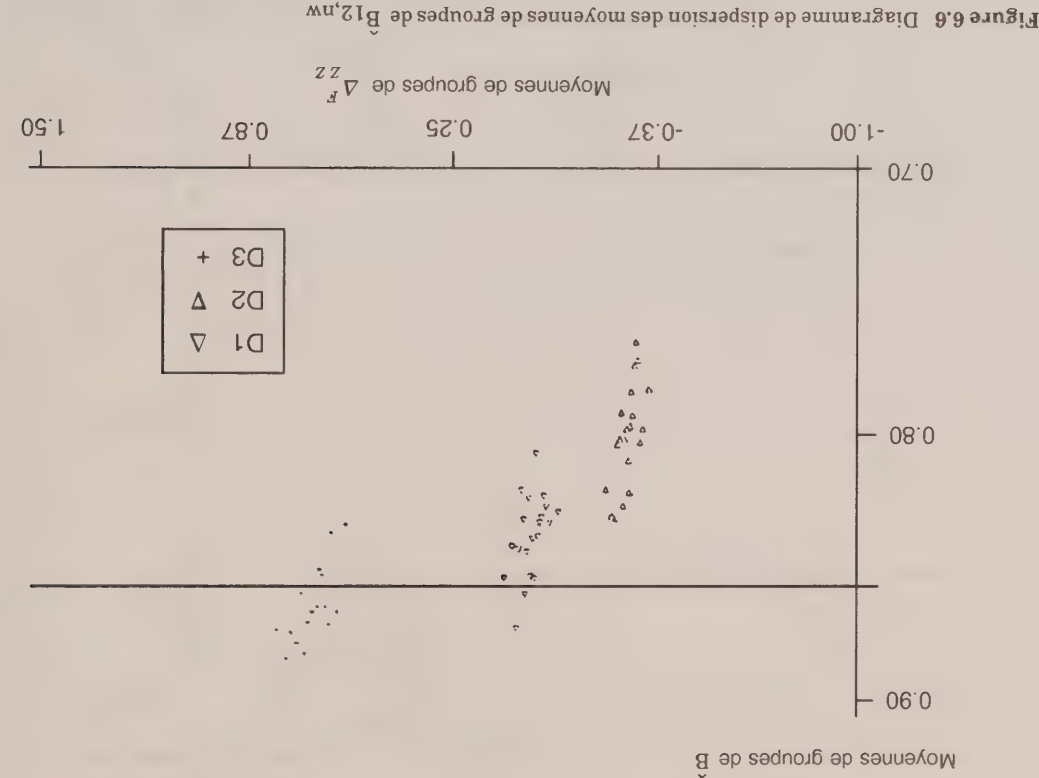


Figure 6.6 Diagramme de dispersion des moyennes de groupes de B12,nw

Les graphiques de l'analyse conditionnelle sont reproduits dans les figures 6.4, 6.5 et 6.6. Nous voyons d'après la figure 6.4 que l'estimateur "ml" est approximativement non biaisé conditionnellement en ce qui concerne les plans D1 et D3 et qu'il n'est pas plus biaisé conditionnellement en ce qui a trait au plan D2. La figure 6.5 nous montre que l'estimateur "pwnl" n'est pas plus biaisé conditionnellement dans l'un ou l'autre des plans. Nous voyons d'après la figure 6.6 que l'estimateur moyen "nw" n'est que légèrement plus biaisé conditionnellement dans chacun des trois plans.

Étude de simulation 3

Echantillonnage répété dans une population "réelle" à plusieurs variables.

Dans cette étude de simulation, nous nous servons des 6,962 données simples réelles tirées de l'enquête sur les dépenses des familles pour constituer la population finie. Nous considérons les mêmes variables que dans la section 3 et tirons successivement des échantillons de cette population pour analyser les propriétés de robustesse des trois estimateurs par régression. Nous nous attendons que la population "réelle" ne respecte aucune des hypothèses de normalité. Les résultats de l'analyse inconditionnelle figurent dans les tableaux 6.9, 6.10 et 6.11. Nous constatons que l'estimateur moyen "nw" est le plus efficace et qu'il est approximativement non biaisé inconditionnellement pour tous les plans d'échantillonnage. L'estimateur "ml" est moins biaisé et plus efficace que l'estimateur "pwnl" en ce qui concerne les plans à probabilités inégales. Les graphiques de l'analyse conditionnelle sont reproduits dans les figures 6.7, 6.8 et 6.9. Nous voyons d'après la figure 6.7 que l'estimateur "ml" est approximativement non biaisé conditionnellement pour les plans D1 et D2 mais qu'il est affecté d'un faible biais conditionnel pour ce qui est du plan D3. La figure 6.8 nous montre que l'estimateur "pwnl" n'est pas plus

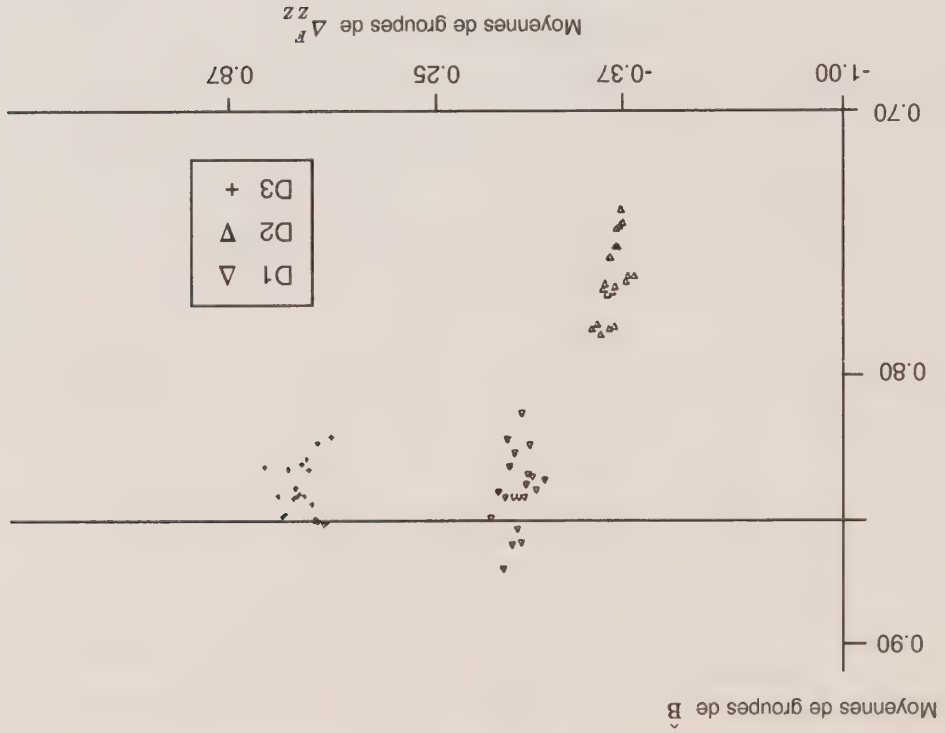


Figure 6.4 Diagramme de dispersion des moyennes de groupes de \hat{B} 12,ml

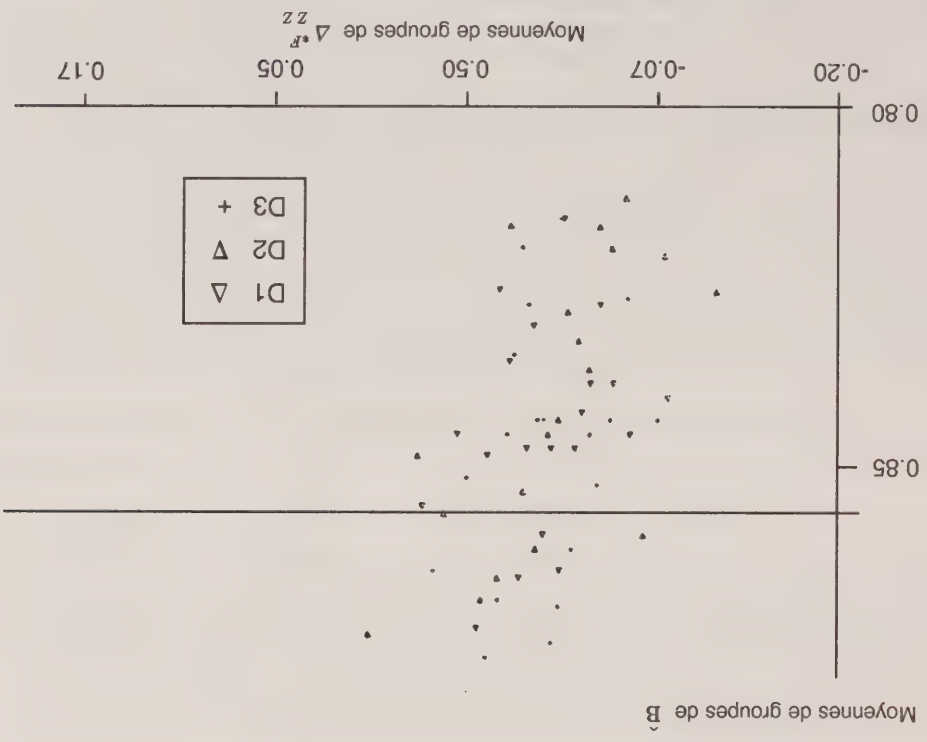


Figure 6.5 Diagramme de dispersion des moyennes de groupes de \hat{B} 12,pwml

Tableau 6.6
 Biais absolus inconditionnels des trois estimateurs de B_{12}
 $N = 6,962, n = 100$ Valeur vraie de $B_{12} = 0.857$

Biais absolus de			
Plan de sondage			
	$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$
D1	0.0119	0.0119	0.0171
D2	0.0923	0.0132	0.5556
D3	0.0124	0.0098	0.0104

Tableau 6.7
 Écart type inconditionnel des trois estimateurs de B_{12}

Écart types			
Plan de sondage			
	$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$
D1	0.0877	0.0877	0.0877
D2	0.0972	0.1230	0.1150
D3	0.0785	0.1110	0.0797

Tableau 6.8

Erreur quadratique moyenne inconditionnelle des trois estimateurs de B_{12}

Erreur quadratique moyenne			
Plan de sondage			
	$B_{12,ml}$	$B_{12,pwml}$	$B_{12,nw}$
D1	0.0078	0.0078	0.0080
D2	0.0180	0.0153	0.0164
D3	0.0063	0.0124	0.0065

Les tableaux 6.6, 6.7 et 6.8 donnent, respectivement, le biais inconditionnel, l'écart-type inconditionnel et l'erreur quadratique moyenne inconditionnelle des trois estimateurs du coef-

ficient de régression.

Nous constatons que l'estimateur " nw " est fortement biaisé et très inefficent en ce qui concerne le plan D2 (répartition croissante), mais qu'il est approximativement non biaisé inconditionnellement et efficace en ce qui a trait aux plans D1 et D3. Comme on pouvait s'y attendre, l'estimateur " $pwml$ " est approximativement non biaisé inconditionnellement pour tous les plans considérés. Bien qu'il soit plus biaisé que " $pwml$ ", l'estimateur " nw " l'est moins que " ml " pour les plans à probabilités inégales. De plus, " nw " est plus efficace que " ml " pour le plan D2 et à peu près aussi efficace pour le plan D3. Il est aussi plus efficace que " $pwml$ " pour le plan D3 (répartition en U).

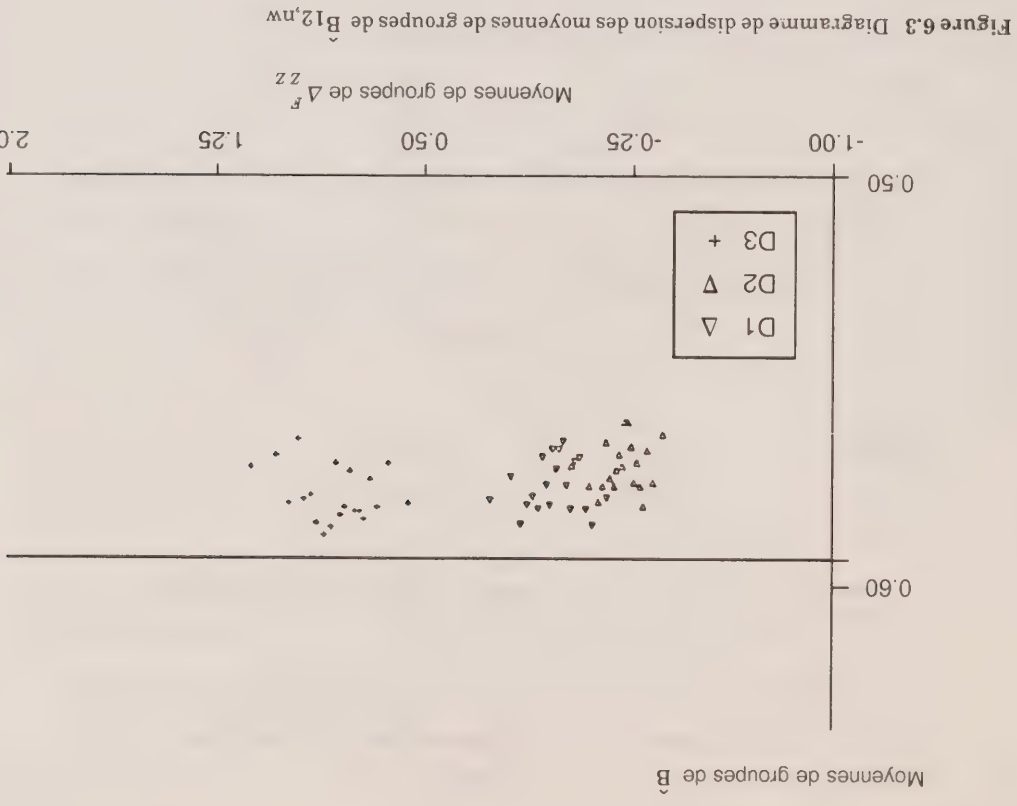


Figure 6.3 Diagramme de dispersion des moyennes de groupes de $B_{12,nw}$

Les diagrammes conditionnels sont reproduits dans les figures 6.1, 6.2 et 6.3. Ils ne révèlent pas d'autres formes de biais que celle que l'on trouve dans le tableau 6.3. Des études antérieures montrent que les biais des estimateurs EAS et des estimateurs simples pondérés en p ont un comportement très régulier (voir Skinner et coll. 1989, chapitres 7 et 8).

Étude de simulation 2

Echantillonnage répété dans une population homogène quadratique

Cette étude de simulation ressemble à une autre effectuée par Holmes (1987). Nous avons généré 6,962 valeurs de (y_{1i}, y_{2i}, z_i) $i = 1 \dots 6,962$, pour population finie en extrayant tout d'abord une valeur z_i de la distribution uniforme $U(0, 10)$. Une fois cette valeur générée, il est possible de calculer les valeurs y_{1i} et y_{2i} correspondantes au moyen des équations suivantes:

$$y_{2i} = m_2 + H_{2z_i} + R_{2z_i^2} + \epsilon_{2i}$$

et

$$y_{1i} = m_1 + H_{1z_i} + R_{1z_i^2} + \epsilon_{1i},$$

où ϵ_{2i} et ϵ_{1i} sont des variables aléatoires tirées de distributions normales de moyenne nulle et de variance constante, et $R_1 \neq 0, R_2 \neq 0$. Comme dans Holmes (1987), nous avons choisi les paramètres de ces équations de telle manière que les régressions de y_1 et de y_2 par rapport à z soient des fonctions monotones croissantes de z et que la régression de y_1 par rapport à y_2 soit approximativement linéaire de manière à faire du coefficient de régression B_{12} un paramètre significatif pour l'estimation.

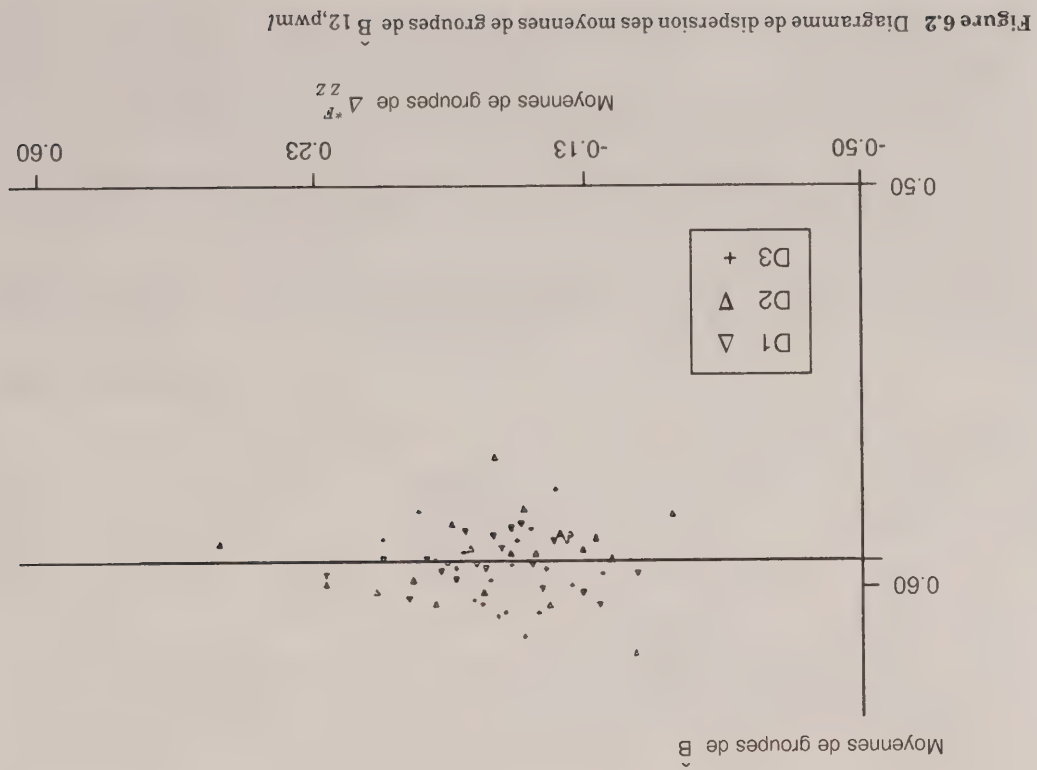
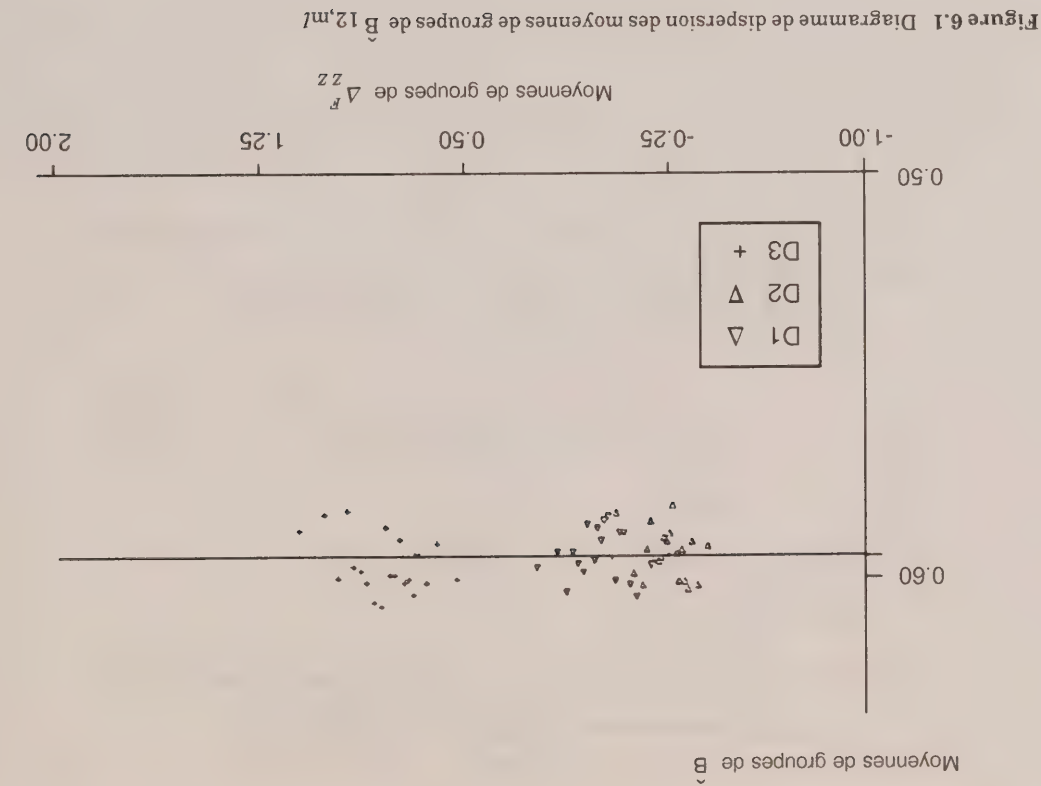


Tableau 6.3
Biais absolus inconditionnels des trois estimateurs de B_{12}
 $N = 6,962, n = 100$ Valeur vraie $B_{12} = 0.595$

Plan de sondage			
$B_{12,m}$		$B_{12,pwml}$	
$B_{12,nw}$		$B_{12,nw}$	
D1	0.0003	0.0003	0.0185
D2	0.0007	0.0019	0.0269
D3	0.0026	0.0018	0.0159

Tableau 6.4

Ecart type inconditionnel des trois estimateurs de B_{12}

Plan de sondage			
$B_{12,m}$		$B_{12,pwml}$	
$B_{12,nw}$		$B_{12,nw}$	
D1	0.0500	0.0500	0.0507
D2	0.0522	0.0693	0.0531
D3	0.0486	0.0710	0.0503

Tableau 6.5

Erreur quadratique moyenne inconditionnelle des trois estimateurs de B_{12}

Plan de sondage			
$B_{12,m}$		$B_{12,pwml}$	
$B_{12,nw}$		$B_{12,nw}$	
D1	0.0025	0.0025	0.0029
D2	0.0027	0.0048	0.0035
D3	0.0024	0.0050	0.0028

Etude de simulation 1

Dans la première étude de simulation, on a "tiré" les 6,962 valeurs pour population finie d'une distribution normale multidimensionnelle; la matrice de corrélation correspondante est celle qui figure dans le tableau 6.1. Ces observations devraient mettre en valeur l'estimateur $B_{12,m}$. Les tableaux 6.3, 6.4 et 6.5 donnent, respectivement, les biais inconditionnels, les écarts-types inconditionnels et l'erreur quadratique moyenne inconditionnelle. Comme prévu, l'estimateur $B_{12,m}$ est celui qui a la meilleure erreur quadratique moyenne. Le nouvel estimateur $B_{12,nw}$ a des qualités surprenantes; son biais est élevé mais son écart-type est comparable à celui de $B_{12,m}$. La valeur du biais est comparable à celles que l'on trouve dans d'autres études pour des populations très lisses (linéaires); voir, par exemple, Casser et Engel (1990). Il faut une très forte largeur de bande pour reproduire une fonction très lisse.

Tableau 6.1
Valeurs de paramètres tirées de la population réelle

Variable	E.T.	Matrice de corrélation
y_1 Dépenses totales	0.668	1
y_2 Revenu total	0.849	0.75
z Dépenses d'alimentation	0.658	0.41
		0.28
		1

Tableau 6.2
Plans d'échantillonnage stratifié

Plans	n_1	n_2	n_3	n_4	n_5	Symbole
D1 Répartition proportionnelle	20	20	20	20	20	Δ
D2 Répartition croissante	5	9	16	30	40	∇
D3 Répartition de U	40	8	4	8	40	+

Les plans d'échantillonnage utilisés s'inspiraient de ceux employés par Holt, Smith et Winter (1980). Désignons un plan d'échantillonnage aléatoire stratifié par $(n_1, \dots, n_h, n_h \text{ unités étant prélevées dans la strate } h, h = 1, \dots, 5)$. Les plans d'échantillonnage ainsi que les symboles utilisés dans les graphiques figurent dans le tableau 6.2.

Pour les divers plans d'échantillonnage stratifié, nous avons prélevé 1,000 échantillons indépendants de taille $n = 100$ dans la population finie. Ces 1,000 échantillons ont servi à estimer la distribution d'échantillonnage des diverses statistiques à l'étude. On obtient les résultats inconditionnels en faisant la moyenne des statistiques à l'étude pour les 1,000 échantillons.

Afin d'évaluer les propriétés conditionnelles des estimateurs, nous avons réparti les 1,000 échantillons en 20 groupes de 50 échantillons chacun, ces groupes étant classés dans un ordre précis, c'est-à-dire suivant un mouvement croissant de la valeur de $\Delta_F^{zz} = (S^{zzs} - S^{zz})/S^{zz}$ pour les estimateurs "nw", et "ml", où

$$S^{zz} = N^{-1} \sum U(z_i - \bar{z}_U)^2, \quad S^{zzs} = n^{-1} \sum_s (z_i - \bar{z}_s)^2,$$

$$\bar{z}_U = N^{-1} \sum U z_i, \quad \bar{z}_s = n^{-1} \sum_s z_i,$$

et de la valeur de $\Delta_F^{*F} = (S^{*F} - S^{zz})/S^{zz}$ pour les estimateurs "pwm", où

$$S^{*F} = \sum_s w_i (z_i - \bar{z}_s)^2, \quad \bar{z}_s = \sum_s w_i z_i, \quad w_i = (N\pi_i)^{-1} \text{ et } \pi_i$$

désigne la probabilité de sélection de l'unité i , de telle sorte que le premier groupe renferme les 50 échantillons auxquels correspondent les valeurs de Δ_F^{*F} (ou de Δ_F^{zz}) les plus basses, et ainsi de suite jusqu'au 20^{ème} groupe, qui contient les 50 échantillons auxquels correspondent les valeurs de Δ_F^{*F} (ou de Δ_F^{zz}) les plus élevées. Nous supposons que la variation de Δ_F^{zz} (ou de Δ_F^{*F}) à l'intérieur de chaque groupe est faible. On peut alors représenter graphiquement la distribution conditionnelle des divers estimateurs, étant donné Δ_F^{zz} (ou Δ_F^{*F}).

Dans les deux premières études de simulation, le biais, l'écart-type et l'erreur quadratique moyenne sont calculés par rapport à la valeur de B^{12U} dans la population finie générée par le modèle. On est donc en mesure de les comparer aux valeurs tirées de la population "réelle" de la troisième étude de simulation.

Nous allons estimer les éléments de $\tilde{\Sigma}_{yy}$ à l'aide des fonctions suivantes:

- (i) l'estimateur corrigé de Pearson de $\tilde{\Sigma}_{yy}$ fondé sur (4.4);
- (ii) la version pondérée en probabilité de (4.4);
- (iii) un estimateur noyau fondé sur (5.14).

Les estimateurs correspondants de B_{12} , ou de son équivalent pour population finie, B_{12U} , sont désignés par $B_{12,ml}$, $B_{12,pwml}$ et $B_{12,nw}$ respectivement. Dans le premier cas, l'indice "ml" signifie que $B_{12,ml}$ est aussi l'estimateur du maximum de vraisemblance (maximum likelihood estimator - MLE) suivant un modèle normal à plusieurs variables. Dans le troisième cas, l'indice "nw" signifie Nadaraya (1964) et Watson (1964). Les deux premiers estimateurs ont été retenus à cause de leur bon rendement dans des études de simulation antérieures; voir Skinner et coll. (1989 chapitre 8).

Nous avons effectué trois études de simulation différentes. Dans la première, nous avons généré une population normale à plusieurs variables dans le but de comparer le rendement du nouvel estimateur avec celui de l'estimateur du maximum de vraisemblance, qui est optimal pour les populations de ce genre. Dans la deuxième étude de simulation, nous avons généré une population homogène quadratique afin de comparer le rendement des estimateurs dans les cas où seule l'hypothèse de linéarité n'est pas respectée. Enfin dans la dernière étude, nous avons comparé les estimateurs dans le cas où la structure de la population est inconnue; autrement dit, nous avons utilisé une population "réelle". Dans toutes ces études de simulation, nous avons effectué des analyses conditionnelles aussi bien que des analyses inconditionnelles. Le premier type d'analyses nous permet de déterminer si un estimateur particulier est efficace pour certains échantillons et moins efficace pour d'autres, tandis que le second type évalue en moyenne l'efficacité d'un estimateur pour l'ensemble des échantillons qui peuvent être produits à l'aide d'un plan particulier.

Le nouvel estimateur utilise le noyau gaussien

$$W_k(z_i, z_j) = c_i \exp \{ - (z_i - z_j)^2 / 2k^2 \}, \quad i \in U, \quad j \in s,$$

où $c_i = 1 / \sum_{j \in s} \exp \{ - (z_i - z_j)^2 / 2k^2 \}$. En opérant une simulation avec différentes valeurs de k , la largeur de bande, nous avons pu constater que l'erreur quadratique moyenne était relativement la même pour un très grand nombre de valeurs de k et que cette constance relative avait été rendue possible grâce à un jeu de réduction du biais et d'accroissement de la variance. Nous avons choisi pour k des valeurs qui produisaient un biais relativement faible pour chaque plan d'échantillonnage stratifié.

Comme la population "réelle" dont nous disposons consistait en 6,962 observations tirées de l'enquête de 1975 sur les dépenses des familles au Royaume-Uni, nous avons construit, pour nos simulations, trois populations ayant cette taille, avec comme vecteur de moyennes et matrice de covariance

$$\tilde{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \tilde{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}.$$

Les valeurs réelles de $\tilde{\Sigma}$ figurent dans le tableau 6.1.

La variable de plan, z , repose sur les dépenses d'alimentation, la variable indépendante sur le revenu total et la variable dépendante sur les dépenses totales. Nous avons divisé la population de 6,962 observations en cinq strates en utilisant la variable de plan comme critère de stratification. L'opération a été effectuée de telle manière que la première strate contenait les 1,393 unités ayant les valeurs z les plus basses, les deuxième, troisième et quatrième strates contenaient 1,392 unités chacune et la cinquième strate contenait les 1,393 unités ayant les valeurs z les plus élevées.

Puisque N est grand, nous proposons d'utiliser la fonction de densité empirique (Parzen 1962) définie par l'équation

(5.11)
$$dF(z) = f(z) = 1/N, \text{ si } z = z_j, j = 1, \dots, N, \\ = 0, \text{ autrement.}$$

En remplaçant $f(z)$ dans (5.10) par son équivalent en (5.11), on obtient l'estimateur

(5.12)
$$\hat{\tilde{\Sigma}}_{yy} = N^{-1} \sum_N^j \hat{h}(z_j).$$

Pour estimer $\tilde{\Sigma}_{yy}$, nous suivons une méthode semblable pour le premier terme du membre de droite de l'équation (5.9). Le second terme, quant à lui, peut s'écrire

(5.13)
$$V^z(\tilde{h}(z)) = \int (\tilde{h}(z) - \tilde{h}_y)(\tilde{h}(z) - \tilde{h}_y)_T^T f(z) dz.$$

Nous proposons comme estimateur

(5.14)
$$V^z(\tilde{h}(z)) = N^{-1} \sum_N^j (\hat{h}(z_j) - \hat{h}_y)(\hat{h}(z_j) - \hat{h}_y)_T^T.$$

Par conséquent, l'estimateur proposé de $\tilde{\Sigma}_{yy}$ est

(5.15)
$$\tilde{\Sigma}_{yy} = N^{-1} \left[\sum_N^j \{ \tilde{\Sigma}_{yy}(z_j) + (\hat{h}(z_j) - \hat{h}_y)(\hat{h}(z_j) - \hat{h}_y)_T^T \} \right].$$

Dans Njenga (1990), on étudie les propriétés asymptotiques de ces estimateurs. Si nous cherchons à estimer $\tilde{\Sigma}_{yy}$, c'est notamment pour effectuer des analyses multidimensionnelles, comme une analyse de régression qui met en rapport au moins deux des composantes de \tilde{y} . Dans la section qui suit, nous donnons les résultats d'une étude de simulation dans laquelle un coefficient de régression simple – régression entre deux variables y est estimé à partir d'échantillons aléatoires stratifiés auxquels correspondent des fractions de sondage différentes.

6. ESTIMATION D'UN COEFFICIENT DE RÉGRESSION: ETUDE DE SIMULATION

Soit $\tilde{y} = (y_1, y_2)^T$, avec comme moyenne $\tilde{\mu}_y = (\mu_1, \mu_2)^T$ et comme matrice de covariance
$$\tilde{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

Nous voulons estimer une fonction de $\tilde{\Sigma}_{yy}$, soit le coefficient de régression linéaire simple

(6.1)
$$B_{12} = \sigma_{12}/\sigma_2^2.$$

On peut estimer ces fonctions paramétriques à l'aide de techniques d'estimation non paramétrique comme les méthodes de lissage linéaire. Parmi les méthodes de ce genre, notons l'estimation-noyau (voir, par ex. Gasser et Müller 1979), la régression locale (voir, par ex. Cleveland 1979) et les fonctions splines de lissage (voir, par ex. Silverman 1985). Nous comptons estimer une à une les fonctions incluses dans (5.1) en nous servant de l'estimateur noyau

$$\hat{\mu}(z) = \sum_{j \in S} W_k(z, z_j) \tilde{y}_j. \tag{5.3}$$

Nous posons la condition selon laquelle la somme des poids doit être égale à un de sorte que l'estimateur soit une moyenne pondérée, et nous utilisons le noyau gaussien, où k est la largeur de bande. Ce genre d'estimateur a fait l'objet de nombreuses études, dont l'une des plus récentes se trouve dans Gasser et Engel (1990).

La structure propre à (5.1) et à (5.2) suppose implicitement que l'on peut écrire

$$\tilde{y}_j = \mu(z_j) + \varepsilon_j, \quad j \in S, \tag{5.4}$$

de sorte que

$$\hat{\varepsilon}_j = \tilde{y}_j - \hat{\mu}(z_j), \quad j \in S. \tag{5.5}$$

Par conséquent,

$$\hat{\varepsilon}_j \hat{\varepsilon}_j^T = (\tilde{y}_j - \hat{\mu}(z_j)) (\tilde{y}_j - \hat{\mu}(z_j))^T \tag{5.6}$$

est un estimateur de $\tilde{\Sigma}^{yy}(z_j)$. En multipliant chaque terme $\sigma^{ab}(z_j)$ de $\tilde{\Sigma}^{yy}(z_j)$ par un facteur d'ajustement linéaire, on obtient

$$\hat{\sigma}^{ab}(z) = \sum_{j \in S} W_h(z, z_j) \hat{\varepsilon}_{ja} \hat{\varepsilon}_{jb}, \tag{5.7}$$

où $W_h(z, z_j)$ est un noyau à largeur de bande h , laquelle est généralement plus élevée que k , qui sert à l'estimation de la moyenne conditionnelle, (5.3).

Pour ce qui a trait aux estimations des moments marginaux, on a recours aux formules standard, à savoir

$$\mu_y = \hat{E}_z(\mu(z)), \tag{5.8}$$

$$\tilde{\Sigma}^{yy} = E_z(\tilde{\Sigma}^{yy}(z)) + V_z(\hat{\mu}(z)). \tag{5.9}$$

Or,

$$\hat{\mu}_y = \int \hat{\mu}(z) f(z) dz,$$

et l'estimateur proposé est

$$\hat{\mu}_y = \int \hat{\mu}(z) f(z) dz. \tag{5.10}$$

Si nous supposons que les distributions de la superpopulation sont des distributions normales multidimensionnelles, alors

(i) $E(\tilde{y} \mid \tilde{z})$ est linéaire en \tilde{z} , et

(iii) $V(\tilde{y} \mid \tilde{z}) = \tilde{K}$, indépendant de \tilde{z} .

Suivant ces hypothèses de linéarité et d'homoscédasticité, il existe un estimateur de la matrice de covariance, $\tilde{\Sigma}_{yy}$, de \tilde{y} basé sur un modèle, qui s'écrit (voir Skinner et coll. 1989, section 6.4)

(4.4)
$$\tilde{\Sigma}_{yy} = \tilde{V}_{yy} + \tilde{b}_{yz} (\tilde{V}_{zz} - \tilde{V}_{zzu}) \tilde{b}_{yz}^T,$$

où \tilde{V}_{yy} , \tilde{V}_{zz} et \tilde{b}_{yz} sont, dans les deux premiers cas, des matrices de covariance de l'échantillon et, dans le troisième cas, une matrice de coefficients de régression pour des données réputées i.i.d. et provenant de la distribution conditionnelle $f(\tilde{y}^U \mid \tilde{z}^U; \tilde{\lambda})$. L'expression (4.4) est appelée estimateur corrigé de Pearson, en hommage à K. Pearson (1903).

Par des études théoriques et empiriques, Pfeffermann et Holmes (1987) et Njenga (1990) ont montré que les inférences faites à partir de (4.4) ne tiennent plus si on s'éloigne des hypothèses de linéarité et d'homoscédasticité. Nathan et Holt (1980) avaient proposé une version pondérée en p de (4.4) dans le but d'accroître la robustesse. Pour obtenir le nouvel estimateur, il suffit de remplacer toutes les sommes pondérées également dans (4.4) par les sommes pondérées en p correspondantes. On obtient ainsi l'estimateur du maximum de vraisemblance pondéré en probabilité (pwm). Les propriétés de cet estimateur ont été étudiées empiriquement et théoriquement dans Holmes (1987), Njenga (1990) et Skinner, Holt et Smith (1989, chapitre 8). On a observé qu'il avait les mêmes propriétés inconditionnelles que d'autres estimateurs pondérés en p , comme l'estimateur de Horvitz-Thompson de $\tilde{\Sigma}_{yy}$, et des propriétés conditionnelles supérieures. Dans l'étude de simulation de la section 6, nous avons choisi l'estimateur pwm pour représenter l'ensemble des estimateurs pondérés en p . Comme la version pondérée en p de \tilde{V}_{zz} dans (4.4) est un estimateur convergent selon le plan de \tilde{Y}^{zzu} , le nouvel estimateur est un estimateur convergent selon le plan de $\tilde{\Sigma}_{yy}$. Nous allons maintenant examiner une nouvelle méthode robuste basée sur un modèle.

5. UN ESTIMATEUR NON PARAMÉTRIQUE FONDÉ SUR LES MOMENTS

Dans cette section, nous tentons de remédier au manque de robustesse d'estimateurs basés sur un modèle comme (4.4), qui dépendent fortement des hypothèses de linéarité et d'homoscédasticité. Si la population finie est constituée d'observations i.i.d. provenant de la superpopulation et que l'attention se porte spécialement sur les paramètres de la superpopulation et que $\tilde{\Sigma}_{yy}$, dans la distribution marginale de \tilde{y} , la méthode que nous utilisons dans les circonstances tient compte du fait que les données d'échantillon sont des observations i.i.d. tirées de la distribution conditionnelle $f(\tilde{y} \mid \tilde{z})$ tandis que les variables de plan, \tilde{z}^U , constituent un échantillon i.i.d. de taille N tiré de la distribution marginale de \tilde{z} . Pour des raisons de simplicité, nous supposons qu'une seule variable de plan a été utilisée, par exemple une mesure de la taille, de sorte que z soit une variable aléatoire scalaire.

Nous supposons que la moyenne et la matrice de covariance conditionnelles de \tilde{y} étant donné z sont des fonctions lisses de z de forme inconnue. Soit

(5.1)
$$E(\tilde{y} \mid z) = \mu(z),$$

(5.2)
$$V(\tilde{y} \mid z) = \tilde{\Sigma}_{yy}(z).$$

comme un ensemble de variables aléatoires i.i.d. (indépendantes et identiquement distribuées), L'applicabilité de cette hypothèse à des populations qui ont une structure – c.-à-d. répartition en grappes ou stratification – est discutable. Dans cet article, nous tenons pour acquis cette applicabilité, du moins si l'on considère des strates largement définies. Compte tenu de cette proposition, un EAS tiré de la population finie est aussi un échantillon i.i.d. de la superpopulation et les inférences peuvent être faites directement de l'échantillon vers la superpopulation. Si l'échantillon n'est pas un EAS mais qu'il est prélevé à l'aide d'un plan $p(s | \tilde{z}_U)$ qui exploite l'information en \tilde{z}_U , il ne s'agit plus d'un échantillon i.i.d. de la superpopulation. C'est ce qu'on appelle le problème de l'échantillonnage, et l'effet de l'échantillonnage doit être pris en considération dans l'inférence finale.

Le modèle de superpopulation définit une hiérarchie:

superpopulation \supset population finie \supset échantillon.

Si la population finie est un ensemble de variables aléatoires i.i.d. tiré de la superpopulation, les paramètres de la population finie, *par ex.* les moyennes, sont liés aux paramètres correspondants de la superpopulation dans les termes suivants:

$$\bar{y}_U = E_m(\bar{y}_U) + O_p(N^{-1/2}). \tag{4.1}$$

Puisque N est habituellement très grand, une inférence sur \bar{y}_U vaudra bien une inférence sur $E_m(\bar{y}_U)$. Les inférences faites sur \bar{y}_U au moyen des poids p rattachés à la règle d'échantillonnage $p(s | \tilde{z}_U)$ sont le fondement de l'inférence analytique basée sur la randomisation. Notons que cette approche dépend fortement de l'hypothèse d'un ensemble de variables aléatoires i.i.d. pour la population finie.

En ce qui concerne les analyses plus complexes, comme l'analyse de régression logistique, on peut se servir de la méthode du pseudo-EMV, mentionnée dans Skinner et coll. (1989, sect. 3.4.4) et Binder (1983), pour définir le paramètre de population finie pertinent de même que l'estimateur de randomisation. Le paramètre de population finie est défini habituellement au moyen d'une équation d'estimation; voir Godambe (1960) et Godambe et Thompson (1986). Comme dans la section 3, les intervalles de confiance reposent sur la distribution inconditionnelle produite par un plan d'échantillonnage aléatoire répété.

L'inférence analytique basée sur un modèle repose sur trois choses: le modèle intégral de la population d'enquête \tilde{y}_U , les variables de plan \tilde{z}_U et la règle d'échantillonnage $p(s | \tilde{z}_U)$, c'est-à-dire

$$f(\tilde{y}_U, \tilde{z}_U, s; \tilde{\lambda}, \tilde{\phi}) = f(\tilde{y}_U | \tilde{z}_U; \tilde{\lambda}) f(\tilde{z}_U; \tilde{\phi}) p(s | \tilde{z}_U). \tag{4.2}$$

Dans le cas de l'échantillonnage aléatoire, le plan d'échantillonnage ne modifie pas la distribution conditionnelle $f(\tilde{y}_U | \tilde{z}_U; \tilde{\lambda})$ mais transforme la distribution marginale de \tilde{z}_U , qui s'écrit $f(\tilde{z}_U; \tilde{\phi})$ avant l'échantillonnage, en

$$g_s(\tilde{z}_U; \tilde{\phi}) = f(\tilde{z}_U; \tilde{\phi}) p(s | \tilde{z}_U) \tag{4.3}$$

après l'échantillonnage. Par conséquent, l'échantillonnage n'influe aucunement sur les inférences portant sur $\tilde{\lambda}$, alors qu'il a de l'effet sur les inférences portant sur ϕ et, donc, sur $\theta = g(\tilde{\lambda}, \tilde{\phi})$, les paramètres de la distribution marginale $f(\tilde{y}_U; \tilde{\theta})$. Dans le cas des inférences portant sur ϕ et θ , les données de l'échantillon ne peuvent passer pour des données d'un EAS tiré du modèle de superpopulation.

Après l'échantillonnage, les possibilités de conciliation sont minces. La base de l'inférence n'est pas du tout la même. Dans un cas il est question d'une distribution inconditionnelle et dans l'autre, d'une distribution conditionnelle. Royall et Cumberland (1981) ont montré de façon concluante comment cela pouvait faire toute la différence. Du même coup, ils ont prouvé le manque de robustesse de quelques-uns des estimateurs classiques de la variance fondés sur un modèle.

La conciliation des deux approches est possible dans le cas, par exemple, d'un plan d'échantillonnage stratifié. Les tenants de la randomisation comme ceux de la modélisation s'entendent pour dire que l'échantillonnage stratifié est robuste, et en ce qui concerne l'EAS à l'intérieur de strates, les inférences basées sur un modèle et celles basées sur la randomisation s'accordent. Ces remarques viennent confirmer l'une des rares observations positives en ce qui concerne les enquêtes par sondage:

Théorème: La stratification est une bonne chose.

Démonstration: Voir Cochran (1977, chap. 5).

La stratification nous permet d'examiner de plus près la question de la robustesse. Si un tenant de la randomisation et un tenant de la modélisation adoptent tous deux le même mode de stratification et le même plan d'EAS à l'intérieur de strates, ils feront les mêmes inférences pour un échantillon donné. Supposons maintenant que par une analyse plus poussée ou l'addition de nouveaux renseignements, on convient qu'il aurait fallu utiliser un niveau de stratification de plus. Quel effet cette constatation a-t-elle sur les inférences respectives du tenant de la randomisation et du tenant de la modélisation? Celui-ci dira que le modèle initial a été mal spécifié et que, par conséquent, les inférences faites à partir de ce modèle sont biaisées. L'estimateur et la variance du modèle initial seront faux. Par contre, le tenant de la randomisation pourra dire que l'information additionnelle est intéressante et utilisable pour une stratification *a posteriori* des résultats initiaux, mais aussi qu'on peut, si c'est nécessaire, ne pas en tenir compte puisque les inférences initiales seraient toujours valables au sens défini en (3.2), la seule conséquence possible étant une perte d'efficacité. D'un côté, on juge que l'inférence initiale n'est pas robuste; de l'autre, la même inférence est vraisemblablement robuste. Lorsqu'on en fait la moyenne pour des échantillons répétés, le biais du tenant de la modélisation se transforme, pour le tenant de la randomisation, en une composante de la variance d'échantillonnage ou en une perte d'efficacité. En conclusion, si les tenants de la randomisation et les tenants de la modélisation partent de la même base, ils n'interprètent pas la même manière les écarts par rapport à cette base. D'un côté, il s'agit d'un biais; de l'autre, il s'agit d'une variance. Peut-on vraiment parler de robustesse dans un cas et de non-robustesse dans l'autre?

4. INFÉRENCE ANALYTIQUE

En inférence analytique, on ne s'intéresse plus à une fonction connue des valeurs \tilde{y}_i , pour la population finie, de sorte que même si $n = N$, l'inférence renferme toujours un certain degré d'incertitude, aussi faible soit-il. Les tests d'hypothèses sont un exemple du genre; l'hypothèse nulle de l'absence de différence n'a pas de signification dans une population finie de taille fixe. Parmi les objets possibles de l'inférence analytique, notons les paramètres $\hat{\lambda}$ et $\hat{\phi}$, dans le modèle (2.2), ou des fonctions de ces paramètres, comme $\hat{\theta}$ dans (2.1). D'autres objets possibles sont les paramètres des populations finies qui ont un rapport bien déterminé avec la population finie donnée, serait-ce par une structure spatiale ou temporelle. On trouvera une analyse récente des méthodes d'inférence analytique dans Skinner et coll. (1989).

L'inférence analytique débute par la spécification du modèle de superpopulation; cette opération vise à définir la nature des liens qui existent entre la population finie et la superpopulation. On pose souvent l'hypothèse que la population finie est extraite d'une superpopulation

Notons qu'avec cette définition de la robustesse, il ne semble pas nécessaire de préciser ce que sont des conditions idéales ou ce que serait une dérogation à ces conditions. L'échantillonnage aléatoire et des estimateurs convergents sont les seules choses requises. Brewer et Särndal (1983) sont clairs là-dessus:

"Les méthodes d'échantillonnage probabiliste sont robustes par définition; comme elles n'utilisent pas de modèle, il est inutile de se demander ce qui arriverait si le modèle s'avérait erroné." (TRADUCTION)

Comment une méthode statistique peut-elle être si robuste?

Parce qu'elle est entièrement sous le contrôle du statisticien; celui-ci ne tente pas d'introduire des "éléments naturels" dans la structure. La distribution de randomisation a une forme connue et ne dépend pas de paramètres inconnus. Il n'est pas nécessaire de faire des inférences sur $p(s | \tilde{z}_U)$. De même, c'est le statisticien qui décide de la base de l'inférence: échantillonnage répété fond sur la règle $p(s | \tilde{z}_U)$. Les règles d'échantillonnage et les estimateurs utilisés peuvent varier selon les statisticiens, mais la procédure qui est représentée par l'expression (3.1) engendre des propriétés de couverture approximativement justes dans chaque cas et est donc robuste. Voilà un exemple de robustesse-critère. Or, il se peut qu'une procédure donnée ne soit pas efficace pour les totaux de certaines variables. Nous avons déjà souligné l'inefficacité notoire de l'estimateur de Horvitz-Thompson dans les cas où la variable d'enquête est corrélée négativement avec la variable de taille. Lorsqu'on essaie de réaliser des conditions de robustesse et d'efficacité parmi un très grand nombre de variables, on en vient souvent à recommander l'utilisation d'un plan d'échantillonnage aléatoire simple stratifié; voir, par exemple, Godambe (1982), Hansen et coll. (1983).

Dans l'inférence basée sur un modèle, le statisticien s'amuse à modéliser la "nature". C'est lui qui choisit des distributions de probabilité comme $f(\tilde{y} | \tilde{z}_U; \tilde{\lambda})$, mais la forme réelle de ces distributions est inconnue, comme le sont d'ailleurs les valeurs des paramètres. Si un estimateur t_s de T , est choisi, son espérance mathématique et sa variance dépendront du choix du modèle. Les écarts par rapport au modèle peuvent amener une modification de la moyenne et de la variance et, par conséquent, une modification des intervalles de confiance qui sont le résultat de l'application du théorème limite central aux résidus du modèle. La robustesse attribuable au théorème limite central est plus limitée dans l'inférence basée sur un modèle que dans l'inférence fondée sur la randomisation puisque dans le premier cas, elle ne s'applique qu'aux résidus. Il est possible d'éliminer certains écarts par rapport au modèle en choisissant un plan de sondage convenable, comme le font Royall et Herson (1973a,b), mais on ne pourra jamais obtenir une robustesse parfaite. La base de l'inférence est aussi totalement différente. Au lieu d'une distribution inconditionnelle fondée sur l'échantillonnage répété, l'inférence basée sur un modèle utilise une distribution conditionnelle fondée sur l'échantillon s . Ces deux approches sont-elles conciliables? Avant l'échantillonnage, c'est-à-dire au moment du choix des méthodes, ces approches peuvent être conciliées. Les tenants de la première position comme ceux de la seconde se servent de la même information préalable, \tilde{z}_U , et les deux groupes utilisent des modèles pour la proposition de plans de sondage et d'estimateurs et déterminent leurs méthodes en fonction de l'erreur quadratique moyenne globale

$$E_m E_p(t_s - T)^2. \tag{3.3}$$

Les tenants de la randomisation posent habituellement une condition, comme la propriété d'être approximativement non biaisé en p , tandis que les tenants de la modélisation peuvent exiger la propriété d'être approximativement non biaisé selon le modèle, et il est possible de concilier les deux positions en choisissant un plan de sondage conçu de telle sorte que l'estimateur non biaisé selon le modèle est aussi non biaisé en p . De cette manière, on exploite au maximum l'expression (2.2) et on tire le meilleur des deux approches.

L'utilisation de l'inférence de randomisation suppose que l'on abandonne certains principes statistiques, comme celui de la vraisemblance, au profit du théorème limite central, car nous prétendons que suivant un échantillonnage aléatoire répété fondé sur la règle $p(s \mid \mathcal{Z}^U)$

(3.1)

$$t_s - T \over \sqrt{V^d_p(t_s)} \sim N(0,1),$$

pour tout t_s qui est approximativement non biais en p pour T , où N et n sont grands mais n/N est faible. Bien qu'on en ait fait la preuve formelle uniquement suivant l'EAS (échantillonnage aléatoire simple) et des plans semblables, des données empiriques montrent que les propriétés de couverture par randomisation des intervalles de confiance à 95% du type

(3.2)

$$t_s \pm 1.96\sqrt{V^d_p(t_s)},$$

où $V^d_p(t_s)$ est un estimateur convergent de $V^d_p(t_s)$, sont approximativement justes sauf dans le cas de plans extrêmes ou de populations hétérogènes.

Sur cette question, Godambe et Thompson (1977) s'expriment dans les termes suivants:

“On peut interpréter l'utilisation de tels intervalles de la façon suivante:
I: Nous sommes passablement sûrs, a priori, que y appartient au sous-ensemble de R^N pour lequel l'intervalle contient $T(y)$ pour 95% de tous les échantillons possibles.

II: Il n'y a aucune raison de croire que les valeurs y de l'échantillon, combinées à toute autre information dont nous pourrions disposer sur la population, modifient de quelque façon que ce soit l'affirmation exprimée en I. Par conséquent, même une fois l'échantillonnage accompli, nous croyons que si on appliquait de façon répétée le plan de sondage à cette population, l'intervalle contiendrait $T(y)$ dans environ 95% des fois.

La robustesse de l'intervalle vient évidemment de ce qu'il suffit de conditions très larges et essentiellement informelles pour que l'interprétation donnée en I et II soit valide.” (TRANSDUCTION)

Hansen et coll. (1983) expriment une opinion très semblable:

“Pour des plans d'échantillonnage probabiliste, les intervalles de confiance calculés, étant donné des échantillons suffisamment grands, sont justes en ce sens que la probabilité (de randomisation) que ces intervalles renferment la valeur à estimer est égale ou supérieure au niveau de confiance théorique et ce, quelle que soit la répartition des caractéristiques entre les éléments de la population d'où est tiré l'échantillon.”

“Pour la plupart, le mot robustesse signifie que les inférences qui sont faites à partir d'un échantillon ne changeront pas si les hypothèses qui ont été posées ne sont pas respectées. En théorie, et même dans les faits, on crée de la robustesse dans les sondages probabilistes en ayant recours à un échantillonnage avec probabilités connues (c.-à-d. la randomisation) et à des estimateurs convergents et en utilisant un échantillon suffisamment grand pour que le théorème limite central s'applique, de sorte que l'on puisse dire que les estimations sont distribuées approximativement selon une loi normale.” (TRANSDUCTION)

préables, \hat{z}^U . Elle ne dépend d'aucun paramètre inconnu ni des données d'enquête, \hat{y}^U . Cela a pour effet de rendre $p(s | \hat{z}^U)$ non informative car il y a moins de renseignements dans $p(s | \hat{z}^U)$ que dans \hat{z}^U à elle seule. C'est ce qui explique que Godambe ait obtenu des résultats négatifs à propos de l'inférence de randomisation.

Par contraste, l'inférence basée sur un modèle dépend uniquement du terme de l'équation (2.2) réservé au modèle, étant donné que $p(s | \hat{z}^U)$ ne contient pas d'information sur \hat{y}_s . Pour faire des inférences prédictives sur \hat{y}_s , on se sert de la distribution conditionnelle, $f(\hat{y}_s | \hat{z}^U; \hat{\lambda})$, qui est indépendante de la distribution de randomisation, $p(s | \hat{z}^U)$. La règle d'échantillonnage est toujours importante au stade de l'élaboration car elle influe sur l'efficacité et la robustesse, mais elle n'a aucune importance au stade de l'inférence. Par ailleurs, l'échantillonnage aléatoire est un moyen sûr de rendre la règle d'échantillonnage non informative, ce qui en fait une méthode d'échantillonnage scientifiquement acceptable. Toutefois, les inférences basées sur un modèle peuvent ne pas être robustes si elles dépendent largement du choix du modèle, comme le démontrent de nombreux auteurs, notamment Hansen et coll. (1983).

Une solution de compromis serait d'utiliser les deux composantes de l'expression (2.2), soit la distribution de modèle et la distribution de randomisation, dans le choix de l'estimateur. Cette solution avait été proposée par Godambe (1955) en vue de renverser les résultats négatifs qu'il avait obtenus. Il proposait d'utiliser comme critère l'espérance de modèle de la variance de randomisation, c'est-à-dire $E^m V^p(t_s)$, où t_s est l'estimateur du total T pour une population finie. Afin de trouver une solution optimale dans une classe particulière de modèles, Godambe imitait le choix de t_s à la classe des estimateurs non biaisés en p . Cette restriction a fait l'objet de nombreuses critiques et plusieurs auteurs, dont Brewer (1979), Särndal (1980), Isaki et Fuller (1982) et Little (1983), ont, par la suite, proposé un assouplissement en reconnaissant la propriété d'être approximativement sans biais comme une sorte d'équivalence de la propriété d'être sans biais. Cela s'exprime le plus souvent comme la propriété d'être asymptotiquement non biaisé selon le plan, laquelle nécessite la construction d'une suite hypothétique de populations finies dont la taille tend vers l'infini. Bien que ce raisonnement mathématique puisse sembler décevant, la proposition selon laquelle les méthodes, qui sont choisies avant le tirage de l'échantillon, devraient dépendre de la moyenne calculée suivant un modèle d'échantillonnage répété est parfaitement acceptable. Le problème se trouve plutôt dans la question suivante: quelle distribution choisir pour faire les inférences une fois que l'échantillon est tiré?

3. ROBUSTESSE

La robustesse est une notion qui n'est pas très bien définie en statistique. Dans l'*Encyclopedia of Statistical Sciences* de Kotz et Johnson (1988), il est écrit:

"(. . .) *une méthode robuste est une méthode qui est efficace non seulement dans des conditions idéales mais aussi dans des conditions autres.*" (TRANSDUCTION)

Et encore, ajoute-t-on, faut-il savoir ce que l'on entend par "conditions autres" et par "efficacité". Avec cette définition très générale en tête, nous allons étudier la robustesse dans l'inférence de randomisation et l'inférence basée sur un modèle pour des totaux de population finie. L'impression générale est que le premier type d'inférence est robuste alors que le second ne l'est pas.

Les résultats négatifs qu'a obtenus Godambe veulent vouloir dire que l'inférence de randomisation est irréalisable en règle générale. C'est sûrement vrai en ce qui concerne les populations hétérogènes, comme celle de Royall ("axe, ass and box of horseshoes"), ou les populations qui comptent un petit nombre de valeurs extrêmes, mais pour ce qui a trait aux populations homogènes, l'expérience montre de façon très évidente que l'inférence de randomisation non seulement est réalisable mais fonctionne dans un sens bien défini.

la matrice des valeurs pour la population finie. Soit un échantillon s , sous-ensemble de U , prélevé selon certaines règles. Nous nous intéressons ici aux règles qui reposent uniquement sur de l'information préalable, \tilde{z}_U , qui existe sur toutes les unités de la population. Soit \tilde{z}_U l'information préalable pour toute la population et $p(s | \tilde{z}_U)$ la règle d'échantillonnage. Puisque la règle ne dépend pas de \tilde{y}_U , elle est non informative. Si $p(s | \tilde{z}_U)$ est une règle d'échantillonnage aléatoire, elle détermine une distribution de probabilité pour \tilde{r} , l'ensemble des échantillons, laquelle distribution est à la base de l'inférence de randomisation. Les données d'échantillon sont représentées par l'expression $d_s = \{(\tilde{y}_i; \tilde{r}_i) : i \in s\}$. Posons y_s comme la matrice des valeurs pour l'échantillon; par conséquent, un estimateur sera une fonction des données, d_s , et de l'information préalable, \tilde{z}_U , qui comprend l'information supplémentaire. Nous désignons par E_p , et V_p l'espérance mathématique et la variance en ce qui regarde la distribution $p(s | \tilde{z}_U)$. Dans une méthode basée sur un modèle, on suppose, en plus, que les valeurs de la population, \tilde{y}_U , sont des variables aléatoires. Un inconvénient majeur de ce genre de méthode est de spécifier un modèle probabiliste paramétrique pour la distribution conjointe de toutes ces variables aléatoires, laquelle distribution doit reposer sur toute l'information préalable, \tilde{y} compris celle sur la structure des unités de la population et les relations entre ces unités. Par conséquent, les modèles doivent refléter une classification hiérarchique (grappes) et une classification par blocs (strates), de même que des corrélations entre les variables. Cette structure risque de devenir si complexe qu'on limite habituellement l'attention aux matrices de moyennes et de covariances. De façon générale, posons $f(\tilde{y}_U | \tilde{z}_U; \tilde{\lambda})$ comme la distribution conditionnelle pour population finie, où $\tilde{\lambda}$ est un vecteur de paramètres inconnus. C'est là une spécification suffisante en ce qui concerne l'inférence prédictive pour des valeurs de population finie comme les totaux. Pour ce qui regarde l'inférence analytique pour des paramètres de la distribution marginale de \tilde{y} , il faut, en plus, spécifier la distribution marginale des valeurs préalables \tilde{z}_U . Désignons cette distribution par $f(\tilde{z}_U; \tilde{\phi})$; alors, la distribution marginale de \tilde{y}_U est

$$(2.1) \quad f(\tilde{y}_U; \tilde{\theta}) = \int f(\tilde{y}_U | \tilde{z}_U; \tilde{\lambda}) f(\tilde{z}_U; \tilde{\phi}) d\tilde{z}_U,$$

où $\tilde{\theta} = g(\tilde{\lambda}, \tilde{\phi})$ est le paramètre analysé.

En appliquant la règle d'échantillonnage à la population, on obtient les données d_s . La distribution conjointe des données, d_s , et des valeurs préalables, \tilde{z}_U , est

$$(2.2) \quad f(d_s, \tilde{z}_U; \tilde{\lambda}, \tilde{\phi}) = p(s | \tilde{z}_U) \int f(\tilde{y}_U | \tilde{z}_U; \tilde{\lambda}) f(\tilde{z}_U; \tilde{\phi}) d\tilde{y}_s \\ = p(s | \tilde{z}_U) f(\tilde{y}_s | \tilde{z}_U; \tilde{\lambda}) f(\tilde{z}_U; \tilde{\phi}),$$

où \tilde{s} , désigne les unités qui ne sont pas dans s . Cette distribution est le fondement d'une méthode d'inférence basée sur un modèle. Nous désignons par E_m , et V_m l'espérance mathématique et la variance en ce qui regarde le modèle.

L'expression (2.2) implique que la personne qui fait les inférences doit connaître parfaitement la règle d'échantillonnage $p(s | \tilde{z}_U)$ comme les valeurs de \tilde{z}_U . Si ce n'était pas le cas, $p(s | \tilde{z}_U)$ pourrait devenir informative par rapport aux valeurs non observées \tilde{y}_s , voir Scott (1977) et Sugden et Smith (1984) et ne pourrait pas, par conséquent, figurer à l'extérieur de l'intégrale de (2.2).

Dans ce contexte, où il est question à la fois d'échantillonnage aléatoire et de modélisation de valeurs, l'inférence de randomisation correspond au cas où les valeurs \tilde{y}_U sont des constantes inconnues et la distribution de modèle devient dégénérée au point \tilde{y}_U . La seule probabilité qui reste est celle en $p(s | \tilde{z}_U)$, et cette distribution pour l'ensemble \tilde{r} de tous les échantillons possibles est à la base de l'inférence de randomisation. Notons que la distribution de randomisation est déterminée entièrement par la règle d'échantillonnage et les valeurs

Méthodes robustes basées sur un modèle pour des enquêtes analytiques

T.M.F. SMITH et E. NJENGA¹

RÉSUMÉ

Dans cet article, on étudie la notion de robustesse appliquée à la randomisation et à l'inférence fondée sur un modèle pour des enquêtes descriptives et analytiques. Le manque de robustesse qui caractérise les méthodes basées sur un modèle peut être compensé en partie par un plan de sondage élaboré avec soin. À partir de méthodes de lissage, les auteurs proposent des méthodes d'analyse robustes basées sur un modèle.

MOTS CLÉS: Enquêtes analytiques; robustesse; méthodes de lissage.

1. INTRODUCTION

Nous allons étudier la notion de robustesse dans l'inférence pour population finie tant du point de vue de la randomisation que du point de vue des modèles. Dans son article pionnier sur une théorie unifiée du sondage des populations finies, Godambe (1955) n'a pas seulement fait la démonstration de son fameux théorème de la non-existence; il a aussi fait des propositions concernant l'inférence robuste pour population finie. Il a proposé un modèle de superpopulation pour les variables unitaires y_i et a laissé à entendre que la stratégie, c'est-à-dire le choix du plan de sondage et de l'estimateur, devrait dépendre de l'espérance de modèle de la variance d'échantillonnage. Il a de plus noté que la propriété d'être sans biais en p était indispensable si l'on voulait appliquer des méthodes optimales. Ces idées ont été développées dans plusieurs articles, notamment ceux de Godambe (1982) et de Godambe et Thompson (1977). Les résultats de ces études tendent à confirmer l'optimalité de l'échantillonnage avec PPT (probabilité proportionnelle à la taille) et de l'estimateur de Horvitz-Thompson (1952). Or, comme l'efficacité de cette approche dans les enquêtes polyvalentes a été démontrée, nous trouvons les conclusions relatives à l'optimalité et à la robustesse moins convaincantes que les conclusions apparemment négatives sur les fondements de l'inférence.

Le manque de robustesse de nombreuses méthodes basées sur un modèle est notoire – voir Hansen et coll. (1983) – et le gros des ouvrages de Royall et des ses collaborateurs, par exemple Royall et Herson (1973a,b), sont consacrés à l'élaboration de méthodes robustes basées sur un modèle. Après avoir passé en revue ces ouvrages, nous proposons une méthode robuste basée sur un modèle pour estimer un grand nombre de statistiques complexes utilisées dans l'analyse multidimensionnelle de données d'enquêtes; cette méthode tient compte des effets de l'échantillonnage. Ce que nous proposons n'est pas une stratégie mais une procédure qui peut servir à l'analyse de données d'enquête après le tirage de l'échantillon.

2. STRUCTURE FORMELLE

Pour étudier la robustesse, il faut tout d'abord définir une structure formelle pour l'inférence pour population finie à la manière de Godambe (1955). Nous considérons donc une population de N unités désignées par l'ensemble $U = \{1, 2, \dots, N\}$. À chaque unité i correspond un vecteur de valeurs, y_i , qui sera calculé sur la base de l'échantillon, et $\tilde{y}_U = (\tilde{y}_1, \dots, \tilde{y}_N)$ désigne

¹ T.M.F. Smith, University of Southampton, United Kingdom; E. Njenga, Kenyatta University, Kenya.

La force de ces observations réside dans leur portée; en effet, elles s'appliquent à n'importe quelle matrice Z de variables explicatives. En particulier, elles s'appliquent à la matrice X des variables explicatives de notre modèle de travail ainsi qu'aux variables explicatives qui auraient été négligées. Naturellement, la faiblesse de ces observations est qu'elles portent sur le processus d'échantillonnage plutôt que sur les résultats de ce processus. Une proportion prévisible des échantillons qui seront tirés seront fortement déséquilibrés par rapport aux variables explicatives connues X . Si l'équilibre en X est indispensable dans une étude particulière, il ne faudra rien laisser au hasard pour sa réalisation (voir l'étude empirique de Royall et Cumberland 1981). Les méthodes d'échantillonnage au sort restreint, qui garantissent l'équilibre en X de l'échantillon prélevé – la "méthode du panier" de Wallenius (1980) par exemple – pourraient être une solution de compromis acceptable.

Il arrive qu'une variable explicative Z dont on ignorait la valeur au moment de l'échantillonnage devienne connue par la suite, comme dans le cas de la stratification *a posteriori*. Si on constate que l'échantillon prélevé est mal équilibré en Z , on peut en conclure que l'échantillonnage probabiliste n'a pas assuré la protection voulue contre le biais suivant $M(X, Z : V)$; s'il est trop tard pour tirer un nouvel échantillon, on doit utiliser un estimateur qui est non biaisé suivant ce modèle afin d'assurer une protection contre le biais. Bref, l'échantillonnage probabiliste ne garantit pas un équilibre approximatif en Z ; il nous permet simplement de dire que cet équilibre est très probable. L'échantillonnage probabiliste permet aussi d'affirmer qu'un échantillon donné est assez bien équilibré, en l'absence de renseignements visant à prouver le contraire. En revanche, il ne doit pas empêcher de constater l'existence d'un déséquilibre lorsque c'est le cas.

Notons que suivant le plan d'échantillonnage probabiliste ci-dessus, l'estimateur $(1/V^{1/2}I)$ $(1/V^{1/2}Y_s)/n$, qui correspond à $T(X : V)$ si V et $V^{1/2}I$ appartiennent tous deux à $\mathcal{M}(X)$ et si s est inclus dans $B(X : V)$, est non biaisé par rapport à la distribution de probabilité issue du plan d'échantillonnage. Toutefois, si l'échantillon prélevé n'est pas équilibré en X (c.-à-d. que si s n'est pas inclus dans $B(X : V)$), le même estimateur n'est pas sans biais suivant $M(X : V)$.

BIBLIOGRAPHIE

GODAMBE, V.P., et JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations – I. *Annals of Mathematical Statistics*, 36, 1707-1723.

ISAKI, C.T., et FULLER, W.A. (1987). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P.S. (1984). A fresh look at bias-robust estimation in a finite population. Dans *Proceedings of the Section Survey Research Methods, American Statistical Association*, 176-178.

PEREIRA, C.A., et RODRIGUES, J. (1983). Robust linear prediction in finite populations. *Revue Internationale de Statistique*, 51, 293-300.

ROYALL, R.M., et HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

ROYALL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 73, 66-77.

SCOTT, A.J., BREWER, K.R.W., et HO, W.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73, 359-361.

TALLIS, G.W. (1986). On the optimality of balanced sampling. *Statistics and Probability*, 4, 141-144.

TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.

WALLENIUS, K.T. (1980). Statistical methods in sole source contract negotiation. *Journal of Undergraduate Mathematics and Applications*, 0, 35-47.

est optimal selon le modèle précité; on a ainsi le meilleur estimateur linéaire sans biais $\sum x_i^{1/2}$ $\sum x_i^{1/2} y_i/n$ et la variance minimum $\{ (\sum x_i^{1/2})^2/n - N\bar{x}\sigma^2$. Celle-ci se compare avanta-
geusement avec la variance de l'estimateur par quotient dans un échantillon équilibré,
possible de conserver l'optimalité de l'échantillon et de l'estimateur (sans que la variance
n'augmente) en soumettant l'échantillon aux conditions supplémentaires suivantes:

$$\sum x_i^{1/2}/n = N \Big/ \sum_N x_i^{1/2}$$
$$\sum x_i^{3/2}/n = \sum_N x_i^{3/2} \Big/ \sum_N x_i^{1/2}.$$

(3)

Avec les conditions (2) et (3), le MELSB prend la forme élémentaire suivante:

$$\sum_N^s x_i^{1/2} \sum_N^s (y_i/x_i^{1/2})/n,$$

qui est, en fait, l'estimateur de Horvitz-Thompson pour un plan d'échantillonnage avec
probabilité proportionnelle à $x_i^{1/2}$.

4. ECHANTILLONNAGE PROBABILISTE

Les résultats de la section 2 sont importants par rapport à une variable explicative non
observée Z. Si Z était connue pour toutes les unités de la population, comme X l'est, nous
pourrions utiliser dès le départ $M(X, Z : V)$ comme modèle de travail et $\bar{f}(X, Z : V)$ comme
estimateur. Mais supposons que nous ignorons l'importance de Z et que nous utilisons le modèle
de travail $M(X : V)$ et l'estimateur $\bar{f}(X : V)$ alors que c'est $M(X, Z : V)$ qui est pertinent.
Dans ces circonstances, nous dirons qu'un échantillon tiré de $B(X : V)$ est "équilibré en X".
Bien que nous puissions choisir un échantillon qui est équilibré en X, rien ne garantit que ce
même échantillon sera équilibré en Z; et s'il ne l'est pas, notre estimateur est biaisé:

$$E(\bar{f}(X : V) - T) = [(1/n)((1/V^{1/2})(1/V^{1/2}Z_s) - 1/Z]\gamma.$$

où γ est le coefficient de Z: $EY = X\beta + Z\gamma$.

On peut assurer une protection contre de tels biais grâce à l'échantillonnage aléatoire. Si
nous recourons à un plan d'échantillonnage aléatoire avec des probabilités de sélection
 $\pi_i = nv_i^{1/2}/1/V^{1/2}, i = 1, 2, \dots, N$, nous aurons une espérance d'équilibre en Z:

$$E\pi_1 V^{-1/2} Z_s/n = 1/Z/1/V^{1/2},$$

l'indice π indiquant que l'espérance est définie par rapport au plan d'échantillonnage aléatoire
et non par rapport à un modèle de prévision. De plus, si notre plan d'échantillonnage est conçu
de telle manière que $\text{var}_\pi(1/V^{-1/2}Z_s/n)$ tend vers zéro lorsque n augmente, on peut réduire
la probabilité de tirer un échantillon fortement déséquilibré, par exemple un échantillon pour
lequel $|1/V^{-1/2}Z_s/n - 1/Z/1/V^{1/2}| > \delta$, en choisissant une taille d'échantillon, n, suffisamment
élevée. Autrement dit, l'échantillonnage probabiliste peut produire un équilibre en Z "en
probabilité".

d'extension, l'estimateur par régression et l'estimateur par quotient. Quant à l'estimateur optimal pour le quatrième modèle, $T(0,1 : x^2) = \sum_i x_i + (N - n)(x_r \sum_i (y_i/nx_i))$, l'estimateur par moyenne de quotients $F_{HT} = Nx \sum_i (y_i/nx_i)$ en tient lieu lorsque la fraction de sondage n/N est faible.

Une manière de trouver une méthode d'échantillonnage et d'estimation pratique suivant l'un ou l'autre de ces quatre modèles de travail est d'utiliser le meilleur estimateur linéaire sans biais selon le modèle tout en garantissant de la robustesse enchoisissant un échantillon pour lequel l'estimateur demeure non biaisé dans des modèles de régression polynomiaux plus généraux. En ce qui concerne les deux premiers modèles, $M(1 : 1)$ et $M(1, 1 : 1)$, nous constatons que cette méthode crée de la robustesse sans que cela se fasse aux dépens de l'efficacité dans le modèle de travail. Suivant les deux modèles, la protection contre le biais nécessite un échantillon équilibré simple (non pondéré); or, ces modèles satisfont les conditions du théorème 2 avec $V = I$, ce qui implique que l'échantillon équilibré simple est optimal.

En ce qui concerne les deux autres modèles toutefois, il est plus difficile de concilier robustesse et efficacité. Nous avons noté dans la section 1 que l'estimateur par quotient était l'estimateur optimal dans $M(0, 1 : x)$, et tandis que l'échantillon optimal est constitué des n unités qui maximisent \bar{x}_s , la protection contre le biais dans $M(1, 1 : x)$ nécessite un échantillon par lequel \bar{x}_s n'est pas maximisée mais posée égale à la moyenne de la population, \bar{x} . Il en va de même du modèle $M(0, 1 : x^2)$ l'échantillon optimal est, là aussi, celui par lequel est maximisée la moyenne de l'échantillon, \bar{x}_s , mais la protection contre le biais dans des modèles de régression polynomiaux nécessite un échantillon "surcompensé", où la moyenne égale $\sum_i x_i^2 / \sum_i x_i$ (Scott, Brewer et Ho 1978).

Suivant les deux derniers modèles, $M(0, 1 : x)$ et $M(0, 1 : x^2)$, il est possible de créer de la robustesse sans trop de perte d'efficacité en utilisant au départ un modèle de travail plus général. Le théorème 2 est utile à cet égard. Prenons tout d'abord le modèle $M(0, 1 : x^2)$. Si nous utilisons $T(0, 1 : x^2)$ dans un échantillon surcompensé, la variance de l'erreur est $\{(Nx)^2/n - \sum_i x_i^2 + \sum_i (x_i - \bar{x}_s)^2\} \sigma^2$. Mais si nous utilisons le modèle de travail plus général $M(0, 1, 1 : x^2)$ et l'estimateur $T(0, 1, 1 : x^2)$, le théorème montre que tout échantillon où $\bar{x}_s = \sum_i x_i^2 / \sum_i x_i$ est optimal, avec, comme variance minimum, $\{(Nx)^2/n - \sum_i x_i^2\} \sigma^2$. Par ailleurs, il est possible d'obtenir une protection contre le biais dans des modèles de régression polynomiaux encore plus généraux sans aucune perte d'efficacité en imposant les contraintes rattachées à la condition $B(X : V)$, c.-à-d. $\sum_i x_i^{j-1} / n = \sum_i^1 x_i^j / \sum_i^1 x_i^j$, $j = 0, 3, \dots, J$. Compte tenu de ces contraintes appliquées à l'échantillon, et que l'on désigne collectivement par l'appellation "équilibré en π ", $T(0, 1, 1 : x^2)$ est l'estimateur par moyenne de quotients (Kott 1984). Cet échantillon et cet estimateur conservent leur optimalité dans tous les modèles de type $M(\delta_0, 1, 1, \delta_3, \dots, \delta_J : x^2)$.

Il n'existe pas toujours des échantillons équilibrés $B(X : V)$. L'exemple ci-dessus le montre; lorsque n augmente à un point tel que $n/N > N(x^2) / \sum_i x_i^2$, il ne peut y avoir d'échantillon équilibré en π car, si c'était le cas, la formule de la variance deviendrait négative. Notons que la condition $n/N > N(x^2) / \sum_i x_i^2$ suppose que $\max(x_i) > Nx/n$, de sorte que dans des populations de ce genre, il n'existe pas de plan d'échantillonnage aléatoire avec probabilité de sélection proportionnelle à x .

Pour généraliser l'autre modèle, $M(0, 1 : x)$, de manière que le théorème s'applique, nous pouvons ajouter une variable explicative, $x_{1/2}$:

$$E(Y_i) = \beta_{1/2} x_{1/2}^i + \beta_1 x_i^i$$
$$\text{var}(Y_i) = \sigma^2 x_i^i$$

Suivant le théorème 2, tout échantillon qui satisfait l'équation

$$\sum_i x_{1/2}^i / n = \sum_i x_i^i / \sum_i x_{1/2}^i$$

$$\begin{aligned} T(X:V) &= (N/n)1^s_X \\ \text{var}(T(X:V)) &= [(N/n) - 1]1^sV1\sigma^2. \end{aligned} \tag{1}$$

Le théorème suivant montre que si $V = I$, la variance en (1) est la plus petite possible, c'est-à-dire que les échantillons équilibrés $B(X: I)$ sont optimaux si $I \in \mathfrak{M}(X)$; ce théorème définit aussi des échantillons optimaux pour une classe de modèles qui ont une structure de variance plus générale.

Théorème 2. Suivant $M(X: V)$, si $V1$ et $V^{1/2}1 \in \mathfrak{M}(X)$, alors

$$\text{var}(T(X: V) - T) \geq [(1/V^{1/2}1)^2/n - 1^sV1]\sigma^2;$$

la variance prend la valeur minimum si et seulement si $s \in B(X: V)$, auquel cas

$$T(X: V) = (1^sV^{1/2}1)(1^sV_s^{-1/2}Y_s)/n.$$

Démonstration: Puisque $V1 \in \mathfrak{M}(X)$, la quantité à minimiser est $a^sA_s^{-1}a$, où $a = X^s1$ (lemme 1). Or, $V^{1/2}1 \in \mathfrak{M}(X)$ suppose qu'il existe un p -vecteur c_1 pour lequel $V^{1/2}1 = Xc_1$ et comme V est diagonale, nous sommes sûrs que $V_s^{1/2}1_s = X_s^sc_1$ pour chaque échantillon s . Il s'ensuit que $c_1^sA_sc_1 = n$ et l'inégalité recherchée découle de l'inégalité de Schwarz:

$$(a^sA_s^{-1}a)(c_1^sA_sc_1) = (a^sA_s^{-1}a) \cdot n \geq (a^sc_1)^2.$$

La condition nécessaire et suffisante pour l'égalité est $a' = kc_1^sA_s$, où $k = 1^sV_s^{-1/2}/n$. Cela équivaut à $s \in B(X: V)$ puisque $c_1^sA_s = 1^s_sV_s^{-1/2}X_s$. On obtient ensuite facilement par des opérations algébriques les expressions simples pour l'estimateur $T(X: V)$ et sa variance.

Les formules du théorème 2 sont bien connues dans la théorie classique des sondages (c.-à-d. celle fondée sur la randomisation). Le MELSB $T(X: V)$ prend la forme élémentaire de l'estimateur de Horvitz-Thompson, $T_{HT} = \sum sY_i/\pi_i$, lorsque π_i , la probabilité de sélection pour l'unité i , est proportionnelle à $v_i^{1/2}$. De plus, la limite supérieure de la variance correspond à celle qu'ont déterminée Godambe et Joshi (1965, théorème 6.1) pour l'espérance de modèle de la variance d'échantillonnage aléatoire.

Supposons que nous avons, pour un modèle de travail $M(X: V)$ qui respecte les conditions du théorème 2, un échantillon optimal s et un MELSB T . Prenons maintenant un modèle plus général $M(X, Z: V)$, qui renferme une ou plusieurs variables explicatives additionnelles Z ; les résultats du théorème 2 sont toujours valables pourvu que l'échantillon appartienne à $B(Z: V)$ aussi bien qu'à $B(X: V)$. L'échantillon et l'estimateur demeurent optimaux suivant le modèle plus général et la variance reste inchangée. Autrement dit, il est possible de conserver l'optimalité dans notre modèle de travail (échantillon à variance minimum et MELSB) et, en même temps, d'assurer une protection contre le biais causé par l'addition de variables explicatives Z en soumettant l'échantillon à la contrainte $B(Z: V)$. Non seulement cette méthode préserve l'estimateur du biais dans $M(X, Z: V)$, mais aussi elle garantit l'optimalité et de l'échantillon et de l'estimateur dans le modèle le plus général. Evidemment, la propriété d'être sans biais est conservée aussi dans le modèle encore plus général $M(X, Z: W)$, où W est une matrice de covariance quelconque.

3. EXEMPLES

Quatre modèles ressortent particulièrement dans la théorie des sondages pour populations finies. D'après la notation utilisée dans la section 1 pour les modèles de régression polynomiaux, ces quatre modèles sont $M(1: 1)$, $M(1, 1: 1)$, $M(0, 1: x)$ et $M(0, 1: x^2)$. Les estimateurs optimaux pour les trois premiers modèles sont, respectivement, l'estimateur avec facteur

2. RÉSULTATS DE BASE

Il est utile, à ce stade-ci, de recourir à l'écriture vectorielle et matricielle, selon laquelle Y est le vecteur de population $(Y_1, Y_2, \dots, Y_N)'$ et le modèle $M(X : V)$ spécifie que $E(Y) = X\beta$ et $\text{var}(Y) = V\sigma^2$, où X est une matrice $N \times p$ de variables explicatives, V est diagonale et le vecteur β et le scalaire σ^2 sont inconnus. Pour un échantillon donné s de n unités, nous inscrivons d'abord les unités échantillonnées, de sorte que

$$Y = \begin{pmatrix} Y_s \\ Y_r \end{pmatrix}, \quad X = \begin{pmatrix} X_s \\ X_r \end{pmatrix}, \quad V = \begin{pmatrix} V_s & 0 \\ 0 & V_r \end{pmatrix},$$

où Y_r est le vecteur de dimension $(N - n)$ correspondant aux unités non échantillonnées, etc. Désignons par 1_s et 1_r les vecteurs $(1, \dots, 1)'$ de magnitude n et $(N - n)$ respectivement. Le total de population est $T = 1_s'Y_s + 1_r'Y_r$. Une fois que l'échantillon s est observé, on connaît le premier terme, $1_s'Y_s$. On obtient le MELSB de T en additionnant à cette quantité le meilleur prédicteur linéaire sans biais de $1_r'Y_r$:

$$\hat{T}(X : V) = 1_s'Y_s + 1_r'X_r\hat{\beta}(X : V),$$

où $\hat{\beta}(X : V) = (X_s'V_s^{-1}X_s)^{-1}X_s'V_s^{-1}Y_s$. La variance d'erreur est

$$\text{var}(\hat{T}(X : V) - T) = 1_r'(X_r'A_s^{-1}X_r + V_r)1_r\sigma^2,$$

où $A_s = X_s'V_s^{-1}X_s$. Ces formules se simplifient lorsque le vecteur V_1 est inclus dans le sous-espace vectoriel créé par les colonnes de X et que nous désignons par $\mathfrak{M}(X)$.

Lemme 1. Si $V_1 \in \mathfrak{M}(X)$, alors

$$\hat{T}(X : V) = 1'X\hat{\beta}(X : V)$$

et, suivant $M(X : V)$,

$$\text{var}(\hat{T}(X : V) - T) = (1'XA_s^{-1}X'1 - 1'V_1)\sigma^2.$$

Démonstration: L'estimateur se simplifie parce que $V_1 \in \mathfrak{M}(X)$ signifie que $V_1 = Xc$ pour un vecteur c quelconque, de sorte que $X_s'1_s = X_s'V_s^{-1}X_sc$, d'où $1_s'X_s\hat{\beta} = c'X_s'V_s^{-1}Y_s = 1_s'Y_s$. La formule de variance découle de la relation $\text{cov}(T, T) = \text{cov}(1'X\hat{\beta}, 1_s'Y_s) = 1'XA_s^{-1}X_s1_s = 1'Xc = 1'V_1$.

Le lemme 1 montre que pour des modèles où $V_1 \in \mathfrak{M}(X)$, l'échantillon n influe sur la variance que par l'intermédiaire de A_s^{-1} . Cela a pour effet de simplifier l'étude de la relation entre la variance et l'échantillon ainsi que la recherche d'échantillons efficaces.

L'ensemble des échantillons qui satisfont

$$1_s'W_s^{-1/2}X_s/n = 1'X/1'W^{1/2}1,$$

où W est une matrice $N \times N$, sera désigné par $B(X : W)$. Lorsque W est la matrice unité, I , $B(X : I)$ désigne l'ensemble des échantillons équilibrés en fonction des colonnes de X . Royall et Herson (1973) ont démontré que, suivant un grand nombre de modèles de régression polynomiaux, les MELSB se simplifient grandement dans des échantillons équilibrés:

Théorème 1. Suivant $M(X : V)$, où $V_1 \in M(X)$, si $s \in B(X : I)$, alors

$$\text{cov}(Y_i, Y_j) = \begin{cases} v_i \sigma^2 & i = j, \\ 0 & \text{autrement} \end{cases}$$

où δ_j est un indicateur qui prend la valeur zéro ou un selon que la variable explicative x_j est incluse ou non dans le modèle. Le meilleur estimateur linéaire sans biais selon ce modèle est désigné par $\bar{T}(\delta_0, \dots, \delta_j : v)$. Notre premier modèle est donc $M(0, 1 : x)$, et $\bar{T}(0, 1 : x)$ est l'estimateur par quotient.

Royall et Herson (1973) ont montré que $\bar{T}(0, 1 : x)$ demeure non biaisé suivant $M(\delta_0, \dots, \delta_j : v)$ pour n'importe quel vecteur $(\delta_0, \dots, \delta_j)$ formé de zéros et de uns, et n'importe quelle série de valeurs v_1, \dots, v_N , si l'échantillon est **équilibré** en x, x^2, \dots, x^J :

$$\sum_N^s x_i^j/n = \sum_1^I x_i^j/n \quad j = 1, 2, \dots, J.$$

Cela signifie que dans un échantillon compensé (ou équilibré), $\bar{T}(0, 1 : x)$ est robuste en ce sens qu'il demeure non biaisé suivant des modèles de régression qui sont beaucoup plus généraux que le modèle de travail $M(0, 1 : x)$. Royall et Herson (1973, section 4.5) ont aussi montré comment, avec un échantillon à peu près équilibré, on peut obtenir assurément un estimateur $\bar{T}(0, 1 : x)$ approximativement sans biais. Ils ont de plus montré que dans un échantillon équilibré, cet estimateur conserve non seulement sa propriété d'être sans biais mais aussi son **optimalité** suivant de nombreux modèles de régression polynomiaux, notamment $M(1 : 1)$, $M(1, 1 : x)$, et $M(0, 1, 1 : x^2)$. Plus particulièrement, l'estimateur est optimal suivant n'importe quel modèle de régression polynomial de degré J ou moins, à la seule condition que la fonction de variance du modèle puisse être exprimée comme une combinaison linéaire des variables explicatives.

Suivant le modèle de travail $M(0, 1 : x)$, on n'assure la robustesse de l'estimateur par quotient dans des échantillons équilibrés qu'au prix d'une grande perte d'efficacité. Suivant ce modèle, l'échantillon pour lequel la variance est minimum est constitué des n unités auxquelles correspondent les valeurs x les plus élevées, et l'efficacité d'un échantillon équilibré est seulement $x/\max_s(x_s)$. (Royall et Herson 1973).

En ce qui concerne l'estimateur linéaire, on obtient des résultats théoriques très comparables aux résultats présentés ci-dessus pour l'estimateur par quotient, exception faite d'une différence majeure. L'estimateur est $\bar{T}(1, 1 : 1) = N[\bar{y}_s + b(\bar{x} - \bar{x}_s)]$, où $b = \sum_s (x_i - \bar{x}_s)y_i/\sum_s (x_i - \bar{x}_s)^2$. Il s'agit de l'estimateur optimal (MELSB) suivant le modèle de régression linéaire à variance constante, $M(1, 1 : 1)$. Lorsque l'échantillon est équilibré, cet estimateur est robuste, conservant sa propriété d'être sans biais (et son optimalité) suivant la même grande catégorie de modèles de régression polynomiaux que pour l'estimateur par quotient. Cependant, à la différence de l'estimateur par quotient, l'estimateur par régression **n'acquiert pas sa robustesse aux dépens de l'efficacité**: la variance selon le modèle de travail $M(1, 1 : 1)$ est réduite au maximum dans des échantillons équilibrés, où $\bar{x}_s = \bar{x}$. Cela s'explique par le fait que la variance d'erreur $E(\bar{T} - T)^2$ est la somme d'une constante et d'un terme proportionnel à $(\bar{x} - \bar{x}_s)^2 \text{var}(b)$. Pour réduire au maximum $\text{var}(b)$, il faut maximiser $\sum_s (x_i - \bar{x}_s)^2$; or, ce terme est supprimé automatiquement dans le cas des échantillons pour lesquels $\bar{x}_s = \bar{x}$.

Existe-t-il d'autres modèles où l'échantillon pour lequel la variance du MELSB est minimum peut aussi assurer une protection contre le biais dans diverses conditions? En particulier, existe-t-il des modèles de ce genre pour des problèmes qui impliquent des fonctions de variance non constante? Nous montrons qu'il existe de tels modèles en posant un théorème qui définit une famille de modèles ayant les propriétés voulues ainsi que les échantillons optimaux correspondants. Les résultats présentés dans cet article intègrent ceux de Kott (1984) et de Tallis (1986) et en sont la généralisation. Ils se rapprochent étroitement aussi des résultats de Pereira et Rodrigues (1983) et de Tam (1986), sans oublier ceux d'Isaki et Fuller (1982.)

Robustesse et optimalité de plan dans des modèles de prédiction pour populations finies

RICHARD M. ROYALL¹

RÉSUMÉ

Dans de nombreux cas de sondage de populations finies, le plan peut être optimal, c'est-à-dire qu'il minimise la variance du meilleur estimateur linéaire sans biais suivant un modèle de travail particulier, mais laissera à désirer au point de vue de la robustesse - l'estimateur aura de fortes chances d'être affecté d'un biais si le modèle de travail est incorrect. Cependant, il existe d'importants modèles selon lesquels le plan de sondage assure efficacité et robustesse. Nous présentons un théorème qui définit ces modèles et les plans optimaux qui s'y rattachent.

MOTS CLÉS: Échantillon équilibré; protection contre le biais; défaillance du modèle; modèle de travail.

1. INTRODUCTION

L'"estimateur par quotient" d'un total de population finie $T = y_1 + \dots + y_N$ est $\bar{T} = N\bar{y}_s/\bar{x}_s$, où $\bar{x} = (x_1 + \dots + x_N)/N$ est la moyenne de population connue d'une variable auxiliaire et \bar{y}_s et \bar{x}_s sont des moyennes d'échantillon. Il s'agit du meilleur estimateur linéaire sans biais (MELSB) de T suivant le modèle M :

$$E(Y_i) = \beta x_i, \\ \text{cov}(Y_i, Y_j) = \begin{cases} \sigma^2 x_i^2 & i = j \\ 0 & \text{autrement.} \end{cases}$$

De façon générale, cet estimateur est biaisé suivant d'autres modèles, qui renferment des fonctions de régression différentes, mais nous verrons plus loin qu'il est possible de s'assurer une protection contre le biais suivant des modèles précis en choisissant soigneusement l'échantillon. Dans cet article, nous nous intéressons particulièrement à des populations auxquelles est censé s'appliquer de façon satisfaisante sinon parfaite un modèle particulier comme M . Nos inférences seront faites à partir de ce modèle. Par exemple, nous qualifierions un estimateur \bar{T} de non biaisé seulement si $E_M(\bar{T} - T) = 0$. Nous reconnaissons, par ailleurs, que ce modèle est une approximation et qu'il pourrait être sérieusement erroné. C'est pourquoi nous le décrivons comme un **modèle de travail**; de plus, nous cherchons des méthodes d'échantillonnage et d'estimation qui soient robustes, c'est-à-dire efficaces, non seulement selon le modèle de travail, mais aussi suivant d'autres modèles qui décriraient mieux les relations entre les variables présentes dans la population étudiée.

Nous désignons par $M(\delta_0, \delta_1, \dots, \delta_J; v)$ le modèle de régression polynomial général:

$$E(Y_i) = \sum_{j=0}^J \delta_j \beta_j x_i^j$$

¹ Richard M. Royall, Johns Hopkins University Baltimore, MD 21205 U.S.A.

qui consiste en des itérations entre deux sous-problèmes qui sont beaucoup moins complexes sur le plan du calcul. Ils présentent des résultats empiriques démontrant que la méthode proposée donne des résultats très satisfaisants.

Couper et Groves étudient si des intervieweurs expérimentés obtiennent des taux de réponse plus élevés que des intervieweurs inexpérimentés, en "neutralisant" les différences dans le plan d'enquête et dans les attributs des populations affectées aux intervieweurs. Après avoir démontré que le rapport est positif et curviligne, ils tentent d'expliquer les mécanismes qui permettent aux intervieweurs expérimentés d'obtenir ces taux de réponse et élaborent sur la nature du rapport. Lahiri et Wang proposent de nouveaux estimateurs du "poids en valeur" et de l'"importance relative" des strates de produits et services, deux paramètres indispensables pour la construction des indices des prix à la consommation aux E.-U. Les estimateurs proposés sont des estimateurs composites qui intègrent de l'information provenant de sources pertinentes. Les auteurs comparent ces estimateurs à quatre autres estimateurs par une étude numérique.

Dans ce numéro

Au mois d'août 1991, l'Université de Waterloo organisait un symposium en l'honneur de monsieur le professeur V. P. Godambe à l'occasion de son 65^{ème} anniversaire de naissance. Les communications qui ont été présentées à ce symposium portaient sur des sujets comme les fondements de l'inférence, la théorie de l'estimation et la théorie des sondages, tous des sujets qui intéressent M. Godambe et que celui-ci a contribué largement à développer. Dédée à M. Godambe, la section spéciale **Inférence à l'aide de données d'enquête** de ce numéro contient quelques-unes des communications du symposium qui ont trait à la théorie des sondages. Collectivement, ces communications traitent de nombreuses questions essentielles relatives à l'inférence fondée sur des données d'enquête, par exemple le rôle de la modélisation, la robustesse, les plans de sondage complexes, les méthodes de rééchantillonnage et les effets de l'imputation.

Royall examine l'estimation basée sur un modèle pour des paramètres de population finie. Il décrit l'opposition entre les plans qui assurent l'efficacité du modèle et ceux qui sont robustes à l'égard d'une défaillance du modèle. La robustesse est concrétisée grâce à des échantillons équilibrés. Royall présente une classe de modèles pour lesquels l'échantillon optimal est déjà équilibré de telle sorte que, pour les modèles de cette classe, robustesse et efficacité sont conciliables. Smith et Njenga étudient l'inférence fondée sur un modèle et celle fondée sur la randomisation pour des enquêtes par sondage et proposent une méthode d'inférence basée sur un modèle qui est robuste et non paramétrique. Par suite de simulations faites avec des données réelles et des données fictives, les auteurs concluent que l'estimateur qu'ils proposent pour l'estimation d'un coefficient de régression est robuste lorsqu'on s'écarte des hypothèses de linéarité et d'homoscédasticité, qu'il est assez efficace et qu'il a des propriétés conditionnelles et inconditionnelles raisonnables.

Rao, Wu et Yue passent en revue les travaux récents sur les méthodes de rééchantillonnage applicables à des plans de sondage complexes, notamment la méthode "jackknife", la méthode BRR (balanced repeated replication) et la méthode "bootstrap". Dans une étude de simulation où ils recourent à une population hypothétique, les auteurs évaluent et comparent des estimateurs de variance et des intervalles de confiance pour la médiane de la population.

Mantel examine le cas de l'estimation par modèle d'une moyenne de population finie établie à partir d'une enquête par sondage. Il propose d'élargir le modèle de telle manière que la moyenne de population finie soit une fonction connue de l'estimation optimale (fondée sur le recensement) d'un paramètre du modèle. Le modèle élargi est alors un compromis entre l'efficacité de modèle et la pertinence de la population finie.

Krieger et Pfeffermann s'intéressent à l'estimation de paramètres de modèle par la méthode du maximum de vraisemblance. Ils décrivent diverses méthodes que l'on trouve dans les ouvrages spécialisés et ils examinent le problème des plans informatifs. Ils proposent l'utilisation de distributions pondérées, où les poids sont définis en modèle comme une fonction des covariables et de la variable étudiée. La méthode proposée donne des résultats assez satisfaisants dans une étude de simulation.

Dans le dernier article de la section spéciale, Särndal examine le problème de l'estimation de la variance lorsque l'imputation est utilisée pour constituer un ensemble complet de données. La variance totale est définie comme la somme de la variance d'échantillonnage et de la variance d'imputation. L'estimateur de variance proposé est un estimateur de l'échantillonnage fondé sur un plan, avec une correction pour biais fondée sur un modèle, en même temps qu'un estimateur de la variance d'imputation fondé sur un modèle. Särndal donne quelques exemples et présente une étude empirique.

Armstrong et Wu exposent le problème de la répartition de l'échantillon pour un plan de sondage général à deux phases comme un problème de programmation non linéaire sous contraintes. En exploitant la structure mathématique du problème, ils proposent une solution

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 18, numéro 2, décembre 1992

TABLe DES MATIÈRES

Dans ce numéro	191
L'inférence avec des données d'enquête	
R. M. ROYALL	
Robustesse et optimalité de plan dans des modèles de prédiction pour populations finies	193
T. M. F. SMITH et E. NJENGA	
Méthodes robustes basées sur un modèle pour des enquêtes analytiques	201
J. N. K. RAO, C. F. J. WU et K. YUE	
Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes	225
H. J. MANTTEL	
Estimation pour population finie à l'aide de fonctions d'estimation	235
A. M. KRIEGER et D. PFEFFERMAN	
Estimation par la méthode du maximum de vraisemblance dans des enquêtes par sondage complexes	241
C. E. SÄRDAL	
Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation	257
J. B. ARMSTRONG et C. F. J. WU	
Une méthode de répartition de l'échantillon pour des plans d'échantillonnage à deux phases	269
M. P. COOPER et R. M. GROVES	
Le rôle de l'intervieweur dans la participation aux enquêtes	279
P. LAHIRI et W. WANG	
Une méthode multivariée pour l'estimation composite des dépenses de consommation en vue du calcul des indices des prix à la consommation aux États-Unis	295
Remerciements	311

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G. J. Brackstone
B. N. Chinnappa

Membres

C. Patrick
D. Roy
F. Mayda (Directeur de la production)
R. Platak (Ancien président)

COMITÉ DE RÉDACTION

Rédacteur en chef

M. P. Singh, *Statistique Canada*

Rédacteurs associés

D. R. Bellhouse, *U. of Western Ontario*

D. Binder, *Statistique Canada*

E. B. Dagum, *Statistique Canada*

J.-C. Deville, *INSEE*

D. Drew, *Statistique Canada*

R. E. Fay, *U.S. Bureau of the Census*

W. A. Fuller, *Iowa State University*

J. F. Gentileman, *Statistique Canada*

M. Gonzalez, *U.S. Office of Management and Budget*

R. M. Groves, *U.S. Bureau of the Census*

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*

Rédacteurs adjoints

P. Lavallée, L. Mach et H. Mantel, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (E.-U.) aux États-Unis, et de 49 \$ (E.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

Techniques d'enquête

Une revue de Statistique Canada

Décembre 1992 Volume 18 Numéro 2



Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1992

Tous droits réservés. Il est interdit de reproduire ou de
transmettre le contenu de la présente publication, sous quelque
forme ou par quelque moyen que ce soit, enregistré ou sur
support magnétique, reproduction électronique, mécanique,
photographique, ou autre, ou de l'emmagasiner dans un système
de recouvrement, sans l'autorisation écrite préalable des
Services de concession des droits de licence, Division de la
commercialisation, Statistique Canada, Ottawa, Ontario, Canada
K1A 0T6.

Décembre 1992

Prix : Canada : 35 \$
États-Unis : 42 \$ US

Autres pays : 49 \$ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa



Statistique
Canada

Statistics
Canada

Canada



374000037

Techniques d'enquête

Une revue de Statistique Canada
Décembre 1992 Volume 18 Numéro 2

Catalogue 12-001

JUN 8 1994

